Group sequential tests for delayed responses

Hampson, Lisa

Award date:
2008

Awarding institution:
University of Bath

Group Sequential Tests for Delayed Responses

submitted by

Lisa Victoria Hampson

for the degree of Doctor of Philosophy

of the

University of Bath

Department of Mathematical Sciences

November 2008

Summary

In practice, patient response is often measured some time after treatment commences. If data are analysed group sequentially, there will be subjects in the pipeline at each interim analysis who have started treatment but have yet to respond. If the stopping rule is satisfied, data will continue to accrue as these pipeline subjects respond. Standard designs stipulate that the overrun data be excluded from any analysis. However, their inclusion may be mandatory if trial results are to be included in a filing for regulatory approval. Methods have been proposed to provide a more complete treatment of the pipeline data, for example Whitehead (1992) and Faldum & Hommel (2007), although several issues remain unresolved.

The work presented in this thesis provides a complete framework for dealing systematically with delayed responses in a group sequential setting. We formulate designs providing a proper treatment of the pipeline data which can be planned ahead of time. Optimal versions are used to assess the benefits for early stopping of group sequential analysis when there is a delay in response. Our new tests still deliver substantial benefits when the delay in response is small. While these fall as the delay increases, incorporating data on a highly correlated short-term endpoint is found to be effective at recouping many of these losses. P-values and confidence intervals for on termination of our delayed response tests are derived. We also extend our methodology to formulate user-friendly error spending tests for delayed responses which can deal with unpredictable sequences of information.

Survival data are a special type of delayed response, where the length of delay is random and of primary interest. Deriving optimal survival tests, we conclude that tests minimising expected sample size for "standard" data are also highly efficient survival trials, achieving a rapid expected time to a conclusion.

Acknowledgements

I would like to take this opportunity to thank my supervisor Prof. Chris Jennison for all his help during the last three years and for introducing me to a subject I have found so fascinating. The work in this thesis has also benefitted from many helpful discussions with Dr. Simon Kirby, whose thoughtful insights into the practical issues associated with group sequential testing have been most useful. Thanks also go to Dr. Kate Hargreaves and Peter Coleman at Pfizer, for keeping up a constant stream of interesting problems during my summer visits to Sandwich.

Finally, I could not have completed this thesis without the encouragement and support of my family, Wouter and all my friends in Bath, in particular everyone in 1West 3.10b. I am also grateful to the EPSRC for their financial support during my studies, to PSI for awarding me the prize for Best Student Poster at their annual 2007 conference and finally to the Society for Clincial Trials for awarding me the Thomas C. Chalmers Student Scholarship Prize for Research Excellence at their 29th Annual Meeting in St Louis, Missouri.

Contents

List of Figures

List of Tables

Glossary

cdf	cumulative distribution function
FDA	Food & Drug Administration (regulatory body to whom new drug applications are made in USA)
GST	group sequential test
mle	maximum likelihood estimate
OBF	O'Brien & Fleming
pdf	probability distribution function
I	information level at an interim analysis
$\tilde{I}$	information level at a decision analysis
I_{max}	maximum information level for a group sequential test
K	number of stages in a group sequential procedure
R	ratio of maximum information level for a group sequential test to that required for a fixed sample test with the same α, β and δ
n_{max}	maximum sample size for a group sequential test
$\tilde{n}$	number of long-term responses available at a decision analysis
t_{max}	time taken to complete recruitment for a group sequential test
t_{final}	time taken to complete a group sequential test
r	delay parameter: the ratio of delay in long-term endpoint to time taken to complete recruitment for a group sequential test
Δ_t	delay in long-term endpoint of direct clinical interest
$\Delta_{1,t}$	delay in short-term endpoint
c	subject accrual rate
τ	correlation between short-term and long-term endpoints
κ	ratio of delay in short-term endpoint to delay in long-term endpoint
p^+	one-sided upper p-value for testing $H_0 : \theta = 0$ against $\theta > 0$
p^-	one-sided lower p-value for testing $H_0 : \theta = 0$ against $\theta < 0$
α	type I error probability
β	type II error probability
δ	alternative at which power of testing procedure is specified
Ω	sample space defined by a testing procedure
Φ	standard normal cumulative distribution function
ϕ	standard normal density function
$\boldsymbol{\beta}$ (boldface)	vector of regression coefficients
$\mathbf{X}^T$	transpose of vector or matrix $\mathbf{X}$
$t_{\nu,\alpha}$	upper α tail point of a Student's t-distribution with ν degrees of freedom
z_α	upper α tail point of a standard normal distribution, i.e., $z_\alpha = \Phi^{-1}(1-\alpha)$
$\mathcal{C}_k$	continuation region at interim analysis k of a group sequential test

Vectors and matrices are indicated by boldface type.

Chapter 1

Introduction

1.1 Clinical trials

A new experimental treatment must pass through several phases of testing before it can be licensed for general use in human subjects. In general, the drug development procedure is often characterised as "learning followed by confirming" (Hung et al., 2006). The early phases of clinical testing are dedicated to exploring the efficacy and safety profiles of the new treatment, while subsequent later phase studies are designed to test the null hypotheses these earlier studies generate. Phase I studies typically involve small numbers of subjects due to the uncertainty inherent in this first stage of testing in humans. Healthy male volunteers are often used, with the exception being oncology; the acute side effects associated with cytotoxic drugs mean that ethics prevent testing in subjects other than those with the condition the treatment is intended to treat. At this first stage of testing, the primary objective is to tentatively explore the safety profile of the drug and identify a range of suitable doses which can be taken forward for further investigation. In Phase II, the focus shifts to exploring the efficacy of the treatment. By the end of this phase, a dose must be selected, and a substantive null hypothesis defined for testing in large-scale Phase III trials. These Phase III trials are designed to confirm the efficacy of the treatment when it is administered as part of normal clinical practice. Efficacy must be confirmed in two independent studies before one can file for regulatory approval for the treatment. While the focus in the later phases of the development process is on investigating efficacy, safety data are monitored continuously throughout.

In modern medicine, randomised controlled trials are regarded as the "gold standard" of clinical testing. They allow us to test the superiority or non-inferiority of a treatment relative to control, and make causal inferences about the effects of a treatment. This

elevated status arises from certain design features. Randomised treatment allocation ensures comparable groups of subjects on each treatment arm, while also eliminating any element of subjectivity from the treatment allocation process. Blinding is used to eliminate bias from the evaluation of the efficacy of the treatment. A trial is said to be double blind if neither the subject nor the clinician knows whether they have been allocated the experimental treatment or control. Blinding subjects equalises the placebo effect while blinding the investigator prevents them "feeling sorry" (Proschan et al., 2006) for, and hence giving preferential treatment to, those subjects they feel have received the inferior treatment.

Suppose we wish to design a clinical trial to make inferences about an unknown parameter θ; we wish to test a null hypothesis H_0 against an alternative H_1. The design is shaped by certain statistical requirements. The probability of a type I error (rejecting H_0 when in fact it is true) must be controlled at some nominal level α, while the test's power (the probability that we reject H_0 when it is false) must be sufficiently high under some alternative value of θ, denoted δ. The simplest of designs is the fixed sample test, where the data can only be analysed once the responses of all recruited subjects have been observed. In practice, investigators often require more complex designs which afford them greater flexibility and efficiency. In response to these practical needs, a rich methodology has been developed for the statistical monitoring of clinical trials with several books being written on this topic, for example, see Whitehead (1997) and Proschan et al. (2006). In this thesis, we are concerned with developing new methods within the group sequential framework, a statistical monitoring approach which is described in more detail in the following section.

1.2 Group sequential methods

1.2.1 Development of group sequential methods

The seminal works of Barnard (1946) and Wald (1947) on sequential methods were developed with the aim of improving the efficiency of sampling methods for munitions testing in World War II. Suppose we observe (without loss of generality) continuous random variables X_i, $i = 1, 2, \ldots$, which are independent and identically distributed according to common probability density function $f(x; \theta)$, the form of which is known up to an unknown constant θ. We wish to test the simple null hypothesis $H_0 : \theta = \theta_0$ against the alternative $H_1 : \theta = \theta_1$. In Wald's sequential probability ratio test (SPRT), after the nth observation is made, the likelihood ratio of the accumulated data,

$$LR(x_1, \ldots, x_n) = \frac{\prod_{i=1}^{n} f(x_i; \theta_1)}{\prod_{i=1}^{n} f(x_i; \theta_0)},$$

is calculated. If $B < LR(x_1, \ldots, x_n) < A$ then the test continues and observation $n+1$ is made. The interval (B, A) is said to be the continuation region of the test. If either $LR(x_1, \ldots, x_n) \geq A$ or $LR(x_1, \ldots, x_n) \leq B$, the test terminates with rejection or acceptance of H_0 respectively. Sampling can continue indefinitely until the sample path of the likelihood ratio exits the continuation region. The critical values A and B are chosen so that the test has approximately overall type I error rate α and power $1-\beta$ at $\theta = \theta_1$.

The sample size on termination of a sequential test is random since the stage at which the test terminates is data dependent. Wald & Wolfowitz (1948) prove that the SPRT is optimal in the sense that among all other sequential procedures with the same error probabilities, the SPRT simultaneously minimises the expected sample size under the null and alternative hypotheses. It is important to note that while the SPRT is optimal in terms of minimising expected sample sizes on termination, the sample size distribution is highly positively skewed, with heavy upper tails. This is because in a worst case scenario sampling can continue indefinitely. Truncated versions of the test where we set a maximum sample size, n_{max}, have been proposed in order to avoid this scenario. By the Neyman-Pearson lemma, the maximum sample size for any sequential test must be strictly greater than n_{fix}, the corresponding fixed sample size. It is not immediately obvious that the Neyman-Pearson lemma should apply to sequential tests. To consider why this is, suppose we observe random variables $X_1, \ldots, X_{n_{fix}}$ which are independently and identically normally distributed with unknown mean θ and known variance σ^2. We wish to conduct a sequential test of $H_0 : \theta \leq 0$ against the one-sided alternative $H_1 : \theta > 0$ with type I error probability α at $\theta = 0$. The sequential test will define a rejection region in the n_{fix}-dimensional sample space for $(X_1, \ldots, X_{n_{fix}})$. Let $\hat{\theta}_k$ denote the maximum likelihood estimator for θ based on the first k observations. Suppose the stopping rule stipulates that we stop early at stage k, $1 \leq k \leq n_{fix}$, for rejection of H_0 if $\hat{\theta}_k \geq c_k$. The rejection region for the sequential test is then given by

$$\mathcal{C}^\star = \bigcup_{k=1}^{n_{fix}} \left\{ (x_1, \ldots, x_{n_{fix}}) : \frac{1}{k} \sum_{i=1}^{k} x_i \geq c_k \right\},$$

even though in practice if a test is terminated at stage k we would not go on to observe $X_{k+1}, \ldots, X_{n_{fix}}$. The fixed sample test stipulates that all n_{fix} observations should be taken. Based on these responses, H_0 is rejected if $\hat{\theta}_{n_{fix}} \geq \Phi^{-1}(1-\alpha)$, where Φ is the cumulative distribution function for a standard normal variate. By the Neyman-Pearson lemma, the critical region in this sample space defined by the fixed sample test is uniformly most powerful, i.e., for a given sample size n_{fix}, the fixed sample test attains the maximum power at any alternative value of θ. Hence, in order for our sequential test to satisfy the same power constraints as the fixed sample test, we must

have $n_{max} > n_{fix}$.

Whilst a sequential test will require more subjects than the usual fixed sample test in a worst case scenario, on average it should offer savings in sample size. These expected savings are one of the primary advantages of incorporating interim monitoring into our testing procedures. These savings in subjects are also accompanied by savings in time. For example, if the new treatment is effective, stopping early for success means that we can jump to the next phase of testing, ultimately reducing the time taken to get the treatment to market. Sequential monitoring is also advantageous for reasons of ethics. If the new treatment is ineffective, stopping the trial for futility and accepting H_0, reduces the number of subjects exposed to an inferior treatment; abandoning a lost cause in this way also enables resources to be diverted to other more promising treatments.

The application of these fully sequential methods to clinical testing was soon made; the first sequential clinical trial to be reported in the medical literature was that of Kilpatrick & Oldham (1954), comparing two types of bronchial dilators. However, the logistical burden associated with analysing the data accumulated after each new observation prohibited the widespread application of sequential methods in practice. The paper of Pocock (1977) heralded the advent of group sequential tests (GSTs) in the form that we now recognise them. In contrast to fully sequential monitoring, GSTs propose that accumulated data only be analysed periodically, after each group of subjects has been observed. GSTs can achieve many of the benefits to be gained from continuous monitoring while imposing a much reduced logistical burden. Eales & Jennison (1992) find that 77% of the savings made by the SPRT on the fixed sample test can be made by GSTs corresponding to a relatively small maximum sample size and only 5 groups of subjects. GSTs are now commonly implemented in practice. Indeed their use is now accepted by regulatory bodies, with the FDA publishing guidelines for incorporating interim analyses in "E9: Statistical Principles for Clinical Trials" in the *Federal Register* of 1998.

1.2.2 Group sequential tests

To illustrate how a GST might proceed in general, suppose that subjects are allocated to either a new treatment or control. Let $X_{A,i}$ and $X_{B,i}$, $i = 1, 2, \ldots$, represent responses of subjects on the new treatment and control respectively. Assume responses of subjects on the new treatment are normally distributed, with unknown mean μ_A and known variance σ^2. Similarly, $X_{B,i} \sim N(\mu_B, \sigma^2)$, $i = 1, 2, \ldots$, and all observations are independent. Define $\theta = \mu_A - \mu_B$ to be the "effect size" for the new treatment. The

first GSTs were formulated to answer the question "does the new treatment differ from control?". Formulating this problem statistically, we wish to test the null hypothesis $H_0 : \theta = 0$ against the two-sided alternative $H_1 : \theta \neq 0$, with type I error rate α and power $1 - \beta$ at $\theta = \pm\delta$. We impose the constraint that a maximum of K groups of $2n$ subjects can be recruited into the study, with K and n pre-specified. Randomisation is blocked so that in each group, n subjects are allocated randomly to each treatment.

After the kth group has been observed, we calculate the standardised test statistic $Z_k = \sqrt{I_k}\hat{\theta}_k$, where $\hat{\theta}_k$ is the maximum likelihood estimate based on all available data at analysis k and $I_k = (2\sigma^2/n)^{-1}$ is Fisher's information for θ at stage k. If $|Z_k| \geq c_k$ we stop with rejection of H_0. Otherwise, if $|Z_k| < c_k$, the test continues with the recruitment of the next group of subjects if $k < K$, and terminates with acceptance of H_0 if $k = K$. The Pocock test corresponds to the case where $c_1 = \ldots = c_K$ and can be thought of as a repeated significance test (Armitage et al., 1969) conducted at a constant significance level, where this level is chosen so that the overall type I error probability is controlled at some pre-specified rate α. Other choices of critical values are possible: the O'Brien & Fleming (1979), Wang & Tsiatis (1987) and Haybittle-Peto (Haybittle (1971) and Peto & Peto (1972)) tests all proceed as in the general case above but with different choices of the constants $\{c_1, \ldots, c_K\}$. In this example of a two-sided GST, early stopping is only permitted for rejection of H_0. However, "inner wedge" tests have been developed, for example by Gould & Pecore (1982) and Pampallona & Tsiatis (1994), which allow early stopping for both rejection and acceptance of H_0.

While two-sided tests can be used to test whether the new treatment is different from control, often in practice we are interested in answering the question "is our new treatment better than control?". In response to this need, one-sided GSTs have been developed which can be used to test the null hypothesis $H_0 : \theta \leq 0$ against the one-sided alternative $H_1 : \theta > 0$, for example the power family of tests of Pampallona & Tsiatis (1994) and the triangular test of Whitehead & Stratton (1983). These tests permit early stopping for either rejection or acceptance of H_0. In contrast to the two-sided tests discussed above which, due to symmetry, are defined by only one set of boundary constants, one-sided tests are defined by two sets of critical values, namely $\{l_1, \ldots, l_K\}$ and $\{u_1, \ldots, u_K\}$ which define the lower and upper boundaries of the continuation region respectively. We set $l_K = u_K$ to ensure that the test terminates properly at the final analysis. In this thesis, we are primarily concerned with finding one-sided tests.

The tests of Pocock, Pampallona & Tsiatis etc. mentioned above are designed under a fixed sequence of group sizes. Error probability constraints will be satisfied only if this sequence is observed when the test is conducted. In practice, however, it may

not be possible to predict group sizes exactly. Error spending tests (Lan & DeMets, 1983) are designed to attain the required type I error rate under any sequence of group sizes. The rationale underlying these tests is that boundary constants are chosen so that at each stage error probabilities are "spent" in response to the group sizes that have been observed subject to the constraint that in the absence of early stopping, the cumulative type I error rate spent at stage K is α. The development of error spending methodology introduced the flexibility needed for GSTs to be implemented as part of normal clinical practice. For a more detailed history of GSTs, the reader is referred to Ghosh (1991).

1.3 Adaptive procedures

1.3.1 Motivation

A recent white paper released by the FDA (FDA, 2004) discussing the problems facing the pharmaceutical industry acknowledged that the drug development process is becoming increasingly costly and inefficient. The cost of getting a drug to market occasionally reaches over $750 million (US) (Jennison & Turnbull, 2005), although the number of licensed drugs is small since development is subject to a high attrition rate. The FDA white paper also reports that a drug starting clinical development has only an 8% chance of eventually being licensed. Many fail due to either a lack of efficacy or safety issues, or a combination of both. "Adaptive" designs, which permit the design of a clinical trial to be modified in response to observed information, have been proposed as one way to improve the efficiency of the development process. When designing a clinical trial, often decisions are made on the basis of limited prior information. The flexibility of adaptive designs is regarded as desirable since it gives one the freedom to modify aspects of the design as more information accumulates.

Cui et al. (1999) cite an instance where it might be pertinent to make such data dependent changes. They describe a trial investigating the efficacy of a new treatment intended to prevent myocardial infarction in patients undergoing coronary bypass graft surgery. Investigators were initially optimistic about the effect of the new treatment, and the sample size was determined to detect a 50% reduction in incidence for the new treatment with 95% power. Halfway through the trial, the data were analysed and the observed reduction in incidence was only half that initially anticipated. However, this was still felt to be a clinically significant effect; investigators applied to increase the sample size to attain 95% power at this observed treatment effect, although this application was rejected due to fear that a data-dependent sample size readjustment of this type would inflate the overall type I error rate. In response to this problem, many

designs have been proposed which allow the sample size to be modified in response to updated estimates of the treatment effect parameter, while controlling the type I error rate at its nominal level. Some designs allow the sample size to be modified in a pre-planned way, for example see Proschan & Hunsberger (1995), while others allow complete flexibility, for example the "variance spending" designs of Fisher (1998) and the designs of Cui et al. (1999).

Bauer & Köhne (1994) propose an adaptive two-stage procedure where aspects of the second stage design, for example the sample size, are allowed to depend on the results from the previous stage in a way that need not be specified ahead of time. At the end of the second stage, a combination test is used to compare the new treatment against control, using the p-values from both stages. Bauer & Köhne propose combining p-values using R.A. Fisher's combination test (1932). However, other methods have been proposed, including combining data according to a weighted inverse normal method. This was first proposed in an adaptive context by L. D. Fisher (1998), where the weights of Z-statistics from each stage are allowed to depend on the results from the previous stages such that the sum of the squares of these weights sums to one. In practice, the general form of the combination function used to combine the data from both stages has to be specified ahead of time. Using this general adaptive two-stage framework, one can derive procedures where at the interim analysis one can adaptively select the patient population in which one intends to demonstrate efficacy (so called enrichment designs), or one can change the primary endpoint used to measure the efficacy of the treatment.

1.3.2 Seamless designs

Typically, a Phase IIb study is designed to compare K doses of an experimental treatment against control. In a more general setting, we may even choose to compare different treatments or different versions of the same treatment. On the basis of this study, supposing at least one treatment looks promising against control, the "best" of these is selected and taken forward to a Phase III study, where it is compared against control. In practice, dose selection is a complex decision involving a trade-off between efficacy and safety since taking forward a dose that is too low means that we run the risk of erroneously rejecting a treatment at the end of Phase III for lack of efficacy, whereas a dose which is too high may cause unacceptable side-effects. Hence, the dose taken forward is not always that associated with the largest estimated treatment effect. At the end of the Phase III study, we conduct a hypothesis test comparing the selected treatment against control. Conventionally, this test is based only on data collected during the Phase III, despite the fact that data have been accumulated on both control and the selected treatment during Phase IIb.

Typically, a Phase IIb study will be followed by a hiatus, during which the Phase III study is planned and regulatory approval sought. If approval is granted, further time will continue to elapse as the logistics of the study are set up. For example, recruitment may be slow to get started and trial investigators must be trained. As a means of reducing this "white space", and ultimately speeding up the development process, seamless designs propose that Phase IIb and III studies are planned up front in one joint protocol. Under this framework, these studies can be regarded as stages in a single two-stage study, where the decision criteria for continuing at the interim analysis are clearly defined ahead of time. Hence, our transition between the phases is seamless. We can think of these seamless designs as treatment selection procedures, since in the simplest case our objective is to select the best out of K experimental treatments and then make comparisons with control. Designs exist which stipulate that the hypothesis test for the selected treatment be based on data collected during both stages. For example, following Bauer & Köhne (1994), Bretz et al. (2006) propose adaptive seamless Phase IIb/III designs, where the p-values from each phase are combined using combination tests. Adjustments are made for the multiple comparisons that are made between the K experimental treatments and control. As of yet, only limited attempts have been made (see Bretz et al. (2006), Jennison & Turnbull (2007)) to quantify the statistical efficiency gains to be made by combining data across the phases in this way.

1.4 Thesis organisation

As outlined in Section 1.2.2, if the stopping rule of a standard GST is satisfied at an interim analysis, we decide whether to reject or accept H_0 on the basis of those data available at the time of the analysis. This is pertinent when response to treatment is immediate or the delay in response is small compared to the accrual period because if we close recruitment at an analysis, the flow of data into the study will halt immediately. Such "immediate responses" will be measured in trials in conditions where an efficacious treatment is expected to produce an effect almost instantaneously. For example, a subject taking painkiller for headache will want pain relief quickly; a pertinent primary endpoint may be a subject's change in pain score one hour after treatment.

In clinical practice however, the trial objective is often to demonstrate the long term benefits of a treatment; inference about the effect size for a new treatment will be based on "delayed responses", i.e., patient responses measured some time after treatment commences. Todd & Stallard (2005) cite an example of a study in cancer-related bone lesions where subjects are followed up for 40 weeks, with the primary endpoint being whether a skeletal event is experienced in this period or not. When testing group

sequentially with delayed responses, at each interim analysis there will exist "pipeline subjects" who have commenced treatment but have yet to be observed. Therefore, the test will overrun if the stopping rule is satisfied at an interim analysis; recruitment will be terminated, but data will continue to accrue as the pipeline subjects are observed.

In Chapters 2 to 9, we develop methodology for dealing with delayed responses in a group sequential setting. Addressing issues of both design and analysis on termination, our methods culminate in an error spending approach which is flexible enough to be implemented in practice. Investigating the properties of optimal versions of our designs, we find the loss of efficiency relative to the immediate response case increases with the delay in response. However, the benefits of early stopping for a rapid time to a conclusion are more robust to increases in delay than the savings in sample size. For example, when delay is approximately 20% of the accrual period, we can still attain more than 80% of the reduction in expected time to a conclusion relative to a fixed sample test seen when response is immediate. However, we can attain around only 50% of the reduction in expected sample size, although in Chapter 7 we show that many of these losses can be recovered from the use of data on a correlated short-term endpoint. Survival, or time-to-event data, is a special type of delayed response where the length of response delay is of primary interest. Optimal GSTs for survival data are derived in Chapter 9 and links between these and tests optimal for more "standard" data explored: we find designs efficient for expected sample size, or information, with normal data achieve a rapid time to a conclusion with survival data. We finish this thesis by discussing the wider implications of our findings and outlining avenues for further work.

Chapter 2

Developing group sequential tests for delayed responses

2.1 A motivating example

Recall from Chapter 1 that we define an immediate response as one that can be observed immediately upon commencement of treatment. Such responses are well handled in a standard group sequential framework. Once the test statistic sample path exits the continuation region defined by the GST, recruitment is closed and the flow of data halts. The stopping rule then directs us whether to reject or accept H_0. When we move from this idealised setting to dealing with delayed responses, which can only be observed once some extended period of observation or treatment has elapsed, problems arise when we try to incorporate interim monitoring into our testing procedure.

To underline the motivation for the work presented in the first part of this thesis, we illustrate some of these problems by means of a simple example. Suppose we have a new experimental treatment which we wish to compare against control. Let $X_{A,i} \sim N(\mu_A, \sigma^2)$, $i = 1, 2 \ldots$, represent the responses of those subjects allocated to the new treatment and $X_{B,i} \sim N(\mu_B, \sigma^2)$, $i = 1, 2, \ldots$, the responses of those on control, where a delay Δ_t is inherent in the response variable. We operate under the simplifying assumptions that all observations are independent and σ^2 is known. Define $\theta = \mu_A - \mu_B$ to be the unknown "effect size" for the new treatment. The objective of the trial is to conduct a $K = 2$ stage group sequential test of $H_0 : \theta \leq 0$ against the one-sided alternative $H_1 : \theta > 0$, with type I error probability α at $\theta = 0$ and power $1 - \beta$ at $\theta = \delta$.

One approach to sequential testing in a delayed response setting might be to apply

standard GSTs using the data at each interim analysis. For instance, in our two-stage testing example we might select a test for H_0 from the existing Δ-family of one-sided tests due to Pampollona & Tsiatis (1994), designed under the assumption that response is immediate. Upon reaching the interim analysis, we find that of the $\tilde{n}_1$ subjects to have been recruited into the trial, only n_1 subject responses have been observed. Hence, we have $\tilde{n}_1 - n_1$ subjects in the pipeline. Suppose that the test statistic based upon all n_1 observed responses is found to have exited the continuation region defined by the boundaries of the GST chosen. Hence, recruitment to the trial is terminated, although "overrun" data will continue to accumulate as the responses of the pipeline subjects are observed. We assume that the pipeline subjects continue to be treated according to protocol. In some cases, ethics will constrain us to wait to observe the responses of all $\tilde{n}_1$ subjects recruited at the time of the interim analysis before deciding whether or not to reject H_0.

Once all of the overrun data have been collected, we face the problem of how they should be incorporated into the decision making procedure. The GST boundaries applied at the interim analysis implicitly assume that a hypothesis decision can be made immediately upon termination of recruitment. Hence, we find ourselves with no obvious decision rule to follow which will determine the hypothesis decision to be made given the final dataset comprised of the n_1 responses observed at the interim analysis and the additional overrun data. If $\tilde{n}_1 - n_1$ is small, the pipeline data are unlikely to change any qualitative conclusions made about θ on the basis of the interim data. However, arguing that rejection of H_0 is still valid if the standardised test statistics based on the interim and final datasets are "similar" in value may not be convincing to a regulator, or do much for the integrity of the trial. Indeed, these arguments become redundant if there are larger differences between the Z-statistics and, fixing n_1, the probability of this increases with $\tilde{n}_1 - n_1$. It is clear that a formal framework for inference is needed to enable a coherent interpretation of the data in such cases.

One solution to the lack of an obvious choice of hypothesis decision rule is to look to methods which have been developed to provide inference on termination of an overrunning GST. For example, Whitehead's (1992) deletion method stipulates that the interim analysis at which the stopping rule is first satisfied be deleted from the records. Suppose the test terminates at stage $k^\star$. Let Z_k denote the standardised test statistic based on the data available at interim analysis k, for $k = 1, \ldots, k^\star - 1$, and let $\tilde{Z}_{k^\star}$ denote the standardised test statistic based on the final dataset at stage $k^\star$. Then, the deletion p-value, p_T, for testing $H_0 : \theta \leq 0$ against $\theta > 0$ upon observing $\tilde{Z}_{k^\star} = z^\star$ is

$$p_T = \mathbb{P}((\ k < k^\star, Z_k \geq u_k\ ) \text{ or } (\ k = k^\star,\ \tilde{Z}_k \geq z^\star); \theta = 0).$$

If recruitment is halted at the first interim analysis, the deletion p-value is the fixed sample p-value based on $\tilde{n}_1$ responses. However, the deletion p-value is not a proper p-value; it does not attain a $U(0,1)$ distribution under $\theta = 0$ and in fact is conservative, claims which are proved later in Chapter 5. Aside from these specific objections, there are also more general concerns about using designs which ignore the structure inherent in our data and then trying to fix the ensuing problems using methods of inference: this seems a somewhat disjointed approach to the question of how one should deal with delayed responses in a sequential setting. A more coherent solution is to recognise at the design stage the delay in the response and design our test accordingly. In the next section, we shall propose a way to do this. We take the expected sample size on termination of our new "delayed response" GSTs to include all subjects recruited. In subsequent chapters, we derive optimal versions of our designs and quantify the reductions on the fixed sample size that are possible for a delayed response.

2.2 Group sequential tests for delayed responses

2.2.1 A new test structure

Consider again the statistical problem described in Section 2.1 where our objective is to design a group sequential one-sided test of $H_0 : \theta \leq 0$ against $H_1 : \theta > 0$ with given error probabilities when there is a delay Δ_t in the subject response. Recruitment is defined to begin at time $t = 0$. Denoting the total number of subjects required by the corresponding fixed sample test by n_{fix}, we impose the constraint that a maximum of $n_{max} = Rn_{fix}$ subjects can be recruited into the trial, where the inflation factor $R > 1$. Subjects will be accrued such that, in the absence of early stopping, all n_{max} subjects will be recruited by time t_{max} and observed by time $t_{final} = t_{max} + \Delta_t$.

From the example given in Section 2.1, it is clear that our new test structure for delayed responses will incorporate two different types of analysis. A K-stage test of this type will require a maximum of $K - 1$ interim analyses to be conducted in the absence of early stopping. At interim analysis $k = 1, \ldots, K-2$, we decide whether to continue to interim analysis $k + 1$ or terminate recruitment. If accrual is halted, we must wait for the pipeline subjects to respond before conducting a decision analysis, at which point H_0 is either rejected or accepted. At interim analysis $K - 1$, we face slightly different choices: we either halt recruitment immediately and follow-up those subjects in the pipeline otherwise we recruit the last group of subjects and wait for all n_{max} subjects to be observed before conducting a decision analysis.

Introducing some notation, let $\hat{\theta}_k$ denote the maximum likelihood estimator for θ based

on all available data at interim analysis k and let $I_k = var^{-1}(\hat{\theta}_k)$ denote Fisher's information for θ. Define $n_{A,k}$ and $n_{B,k}$ to be the numbers of responses observed at interim analysis k on the experimental and control treatments respectively. For a two treatment comparison,

$$I_k = \left(\frac{\sigma^2}{n_{A,k}} + \frac{\sigma^2}{n_{B,k}} \right)^{-1}, \quad k = 1, \ldots, K-1. \tag{2.1}$$

For each $k = 1, \ldots, K-1$, let $Z_k = \sqrt{I_k}\hat{\theta}_k$ be the standardised (or Wald) statistic for testing $H_0 : \theta \leq 0$ at interim analysis k. Similarly, define $\tilde{n}_{A,k}$ and $\tilde{n}_{B,k}$ to be the number of subjects recruited and also the number of responses observed on the experimental and control treatments at decision analysis k. Let $\tilde{\theta}_k$ denote the maximum likelihood estimator for θ based on the data available at decision analysis k and let $\tilde{I}_k = var^{-1}(\tilde{\theta}_k)$ denote Fisher's information for θ. Define $\tilde{Z}_k = \sqrt{\tilde{I}_k}\tilde{\theta}_k$ to be the standardised test statistic for testing $H_0 : \theta \leq 0$ at decision analysis k. Building upon the above framework, we define a K stage one-sided delayed response GST of $H_0 : \theta \leq 0$ against $H_1 : \theta > 0$ as shown below:

At interim analysis $k = 1, \ldots, K-2$

if $l_k < Z_k < u_k$	continue to interim analysis $k+1$,
otherwise	stop accrual, continue to decision analysis k.

At interim analysis $K-1$

if $l_{K-1} < Z_{K-1} < u_{K-1}$	continue to decision analysis K,
otherwise	stop accrual, continue to decision analysis $K-1$.

At decision analysis $k = 1, \ldots, K$

if $\tilde{Z}_k \geq c_k$	reject H_0,	(2.2)
otherwise	accept H_0.	

This new test structure is illustrated in Figure 2-1. One can see that the test is defined by three sets of critical values: the sets $\{l_1, \ldots, l_{K-1}\}$ and $\{u_1, \ldots, u_{K-1}\}$ define the continuation region at each interim analysis, while the decision constants $\{c_1, \ldots, c_K\}$ define the criterion according to which we make our final hypothesis decision. In Section 2.2.2, we shall give formulae for calculating the error probabilities of tests of this form.

The above test structure remains valid under any fixed sequence of information

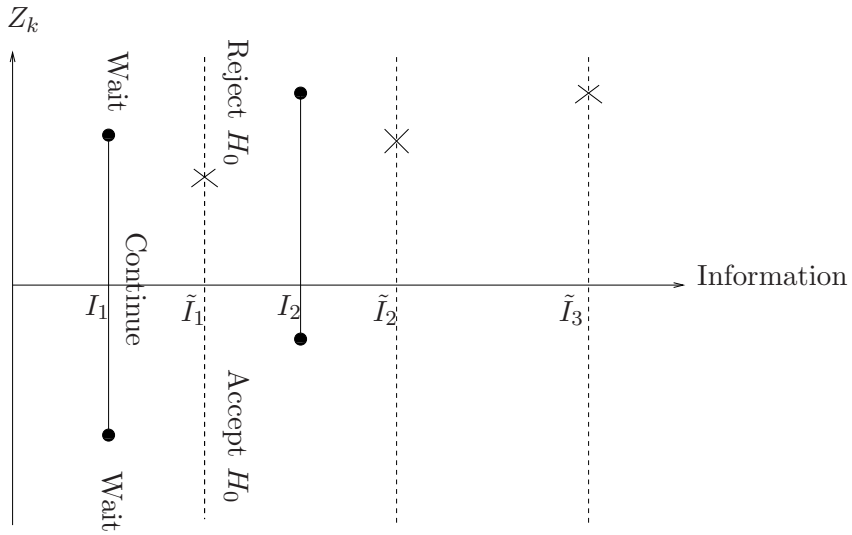


Figure 2-1: Illustration of a three-stage one-sided test of $H_0 : \theta \leq 0$ of the form (2.2). Interim analyses $k = 1, 2$ are scheduled at information levels I_k. If recruitment is stopped early at interim analysis k, we continue to a decision analysis at information level $\tilde{I}_k$, otherwise we continue to a final decision analysis at information level $\tilde{I}_3 = RI_{fix}$.

levels. However, we derive efficiency results for delayed response GSTs under several simplifying assumptions. For a general K-stage test, let t_k and $\tilde{t}_k$ denote the timings of interim and decision analysis k respectively. In the following work, we suppose that the $K-1$ interim analyses are equally spaced in time between Δ_t and t_{max}, so that

$$t_k = \Delta_t + \frac{k}{K}(t_{max} - \Delta_t), \qquad \text{for } k = 1, \ldots, K-1.$$

Nothing can be gained from conducting an interim analysis in the interval $[t_{max}, t_{final}]$ since recruitment has already been terminated by this point and we are now constrained to wait to observe all n_{max} recruited subjects. In addition, following, for example, Galbraith & Marschner (2003), we make the simplifying assumption that recruitment occurs at a constant rate, c. The number of subjects in the pipeline at an interim analysis will be the product of this accrual rate and the delay in the subject response, Δ_t. Note that a long delay and slow accrual rate may result in the same number of pipeline subjects as a short delay and fast accrual rate. Define the delay parameter $r = \Delta_t / t_{max}$ to index the fraction of the test's maximum sample size which is in the pipeline at an interim analysis. For interim analyses equally spaced between Δ_t and t_{max}, we can write

$$t_k = t_{max}\left[r + \frac{k(1-r)}{K}\right], \quad \tilde{t}_k = t_k + r t_{max}, \quad \text{for } k = 1, \ldots, K-1. \tag{2.3}$$

Adapting the analysis timings to the delay in the data helps the efficiency of the test by ensuring at an interim analysis there are still numbers to recruit and hence to save

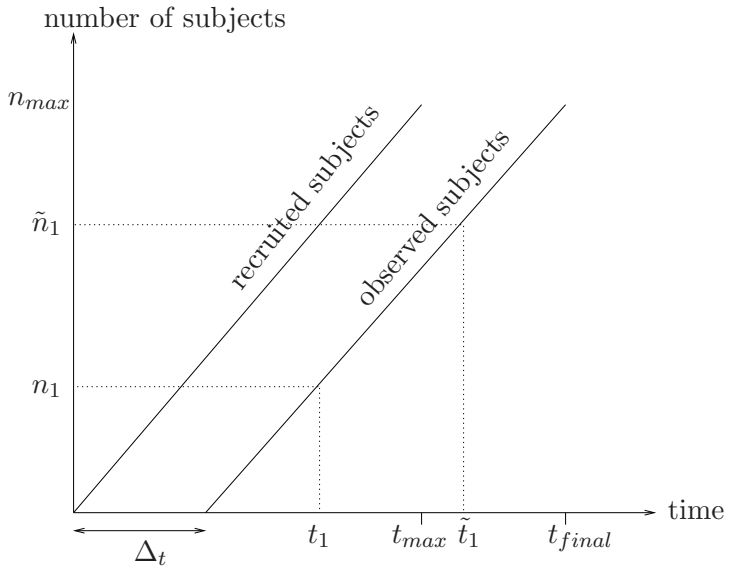


Figure 2-2: How a two-stage delayed response GST might progress in calendar time when there is a delay Δ_t in the response. Analyses are scheduled in time according to (2.3). In the absence of early stopping, all n_{max} subjects will be observed at time t_{final}.

by stopping. Fixing t_{max}, r increases with Δ_t and the interim analyses become more closely spaced in time. If r is close to 1, the number of subjects recruited at any interim analysis will be close to n_{max}. Figure 2-2 illustrates how a two-stage test would progress in calendar time according to this scheduling.

From equation (2.1), we see I_k depends not only on the total number of subjects observed at stage k but also on the proportion of subjects allocated to each treatment, with the most efficient approach being equal allocation. A similar observation holds for $\tilde{I}_k$. In order to eliminate the effect of any imbalance in the allocation ratios on the information for θ, we assume for now that at any time t, equal numbers of subjects will be on each treatment arm. Let n_k and $\tilde{n}_k$ denote the total number of subjects whose responses are observed at interim analysis k and decision analysis k, respectively. Under the assumption of equal allocation, information will be a linear function of sample size:

$$I_k = \frac{n_k}{4\sigma^2} \quad \text{and} \quad \tilde{I}_k = \frac{\tilde{n}_k}{4\sigma^2}. \tag{2.4}$$

Setting a maximum sample size of $n_{max} = Rn_{fix}$ is equivalent to setting a maximum information level $I_{max} = RI_{fix}$, where I_{fix} denotes the fixed sample information level. The number of subjects responses observed at time t is given by $n(t) = c(t - \Delta_t)$. Hence, the information for θ at time t can be written as

$$I(t) = \frac{n(t)}{4\sigma^2} = \frac{c(t - \Delta_t)}{4\sigma^2}.$$

Substituting the analysis timings of (2.3) into this formula, we find that this scheduling corresponds to interim analyses which are equally spaced in information, with rI_{max} units of information in the pipeline at each interim analysis:

$$I_k = \frac{k}{K}(1-r)I_{max} \quad \text{and} \quad \tilde{I}_k = I_k + rI_{max}.$$

We see that the delay parameter $r = \Delta_t/t_{max}$ indexes the amount of information in the pipeline at an interim analysis and hence can be thought of as representing the "real" delay in the system.

One can see that there exist many sets of critical values defining a delayed response GST satisfying given error probability constraints. Therefore, our choice of test for H_0 must be based upon consideration of other properties. In particular, one may seek out tests which are most efficient, or optimal, in some sense. In Chapter 3 we discuss several criteria which can be used to measure a test's efficiency and present methodology which can be used to find tests optimal for these criteria.

2.2.2 Distributional results

In Section 2.2.1 we chose to define the new delayed response GST in terms of standardised test statistics. Alternatively, our test of H_0 could be expressed in terms of the score statistics $S_k = \sqrt{I_k}Z_k$, $k = 1, \ldots, K-1$, and $\tilde{S}_k = \sqrt{\tilde{I}_k}\tilde{Z}_k$, $k = 1, \ldots, K$. For convenience, when deriving distributional results for delayed response GSTs in this section, we choose to work on the score statistic scale. The sequence of information levels upon which our test is based will depend upon r, the delay which is inherent in our system. When r is small, the information in the pipeline at an interim analysis, $\tilde{I}_1 - I_1$, will be small as a fraction of I_1. Hence, the information levels for each stage will be non-overlapping, with $I_k < \tilde{I}_k < I_{k+1}$ for $k = 1, \ldots, K-2$. For r sufficiently large however, this will no longer hold and we find ourselves in the situation where the number of subjects in the pipeline at the first interim analysis is larger than the number of additional responses that will be observed by the next interim analysis should we choose to continue sampling. Under the timing schedule presented in (2.3), we have $\tilde{I}_k > I_{k+1}$ when $r > 1/(K+1)$. Let $I_{(i)}$ denote the ith smallest information level in the sequence $\{I_1, \tilde{I}_1, \ldots, I_{K-1}, \tilde{I}_{K-1}, \tilde{I}_K\}$ and denote by X_i the associated score statistic which is either S_k or $\tilde{S}_k$ for some k. Under this ordering, $X_1 = S_1$ and $X_{2K-1} = \tilde{S}_K$ but X_2 to X_{2K-2} can be various permutations of the remaining S_ks and $\tilde{S}_k$s. Figure 2-3 shows one possible sequence of overlapping information levels and their corresponding ordering when $K = 3$.

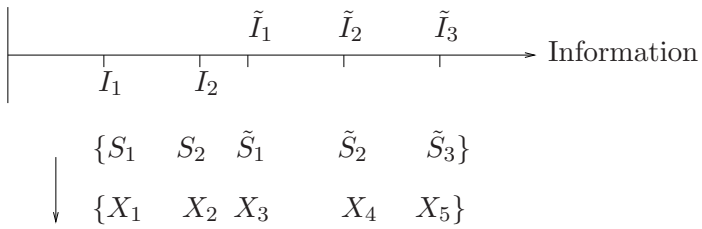


Figure 2-3: A sequence of information levels which may be generated by a three-stage delayed response GST. The score statistics generated by the test are ordered according to their corresponding information levels, where X_i denotes the score statistic based on the ith smallest information level.

Let $\mathcal{F}_i$ denote the set of subject responses upon which X_i is based. Then, $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots \subset \mathcal{F}_{2K-1}$. To illustrate why this should be, consider the scenario depicted in Figure 2-3 where $\tilde{I}_1 > I_2$ and hence $X_2 = S_2$ and $X_3 = \tilde{S}_1$. Should we choose to continue recruitment at interim analysis $k = 1$, the subjects observed in the second group will be the first $|\mathcal{F}_2| - |\mathcal{F}_1|$ pipeline subjects that would be observed if recruitment was closed. Hence, we have $\mathcal{F}_3 \supset \mathcal{F}_2$. We conclude that $X_1, \ldots, X_{2K-1}$ are based on nested subsets of the final dataset available at decision analysis K. Hence, we can extend to our case the results of Jennison & Turnbull (1997, Theorem 1) for the joint distribution of a sequence of test statistics generated by an accumulating body of data. We obtain that in general, when data are normally distributed, for a given sequence of information levels $\{I_k, \tilde{I}_k\}$ the sequence of statistics $\{X_1, \ldots, X_{2K-1}\}$ will have canonical distribution

$$\begin{aligned} &\text{(i)} \quad (X_1, \ldots, X_{2K-1}) \text{ are jointly multivariate normal,} \\ &\text{(ii)} \quad X_k \sim N(\theta I_{(k)}, I_{(k)}), \quad \text{for } k = 1, \ldots, 2K-1, \\ &\text{(iii)} \quad Cov(X_{k_1}, X_{k_2}) = I_{(k_1)}, \quad \text{for } 1 \leq k_1 \leq k_2 \leq 2K-1. \end{aligned} \tag{2.5}$$

This distribution will be correct asymptotically for other data types. It is evident from (2.5) that the increments $X_1, X_2 - X_1, \ldots, X_{2K-1} - X_{2K-2}$ are independent and hence both the sequence $X_1, \ldots, X_{2K-1}$ and the sequence of ordered standardised test statistics are Markov. We can think of the ordered score statistics generated by the delayed response GST as a Brownian motion with drift θ observed at information times $I_{(1)}, \ldots, I_{(2K-1)}$.

A test's error rates can be written as the sum of probabilities for the sequence $(X_1, \ldots, X_{2K-1})$. For example, consider the three-stage GST depicted in Figure 2-3. For $k = 1, 2$, define $\mathcal{C}_k = (l_k, u_k)$ to be the continuation region at interim analysis k. Then, the event that we stop at the second stage with rejection of H_0 can be written as $\{X_1 \in \mathcal{C}_1, X_2 \notin \mathcal{C}_2, X_3 \in \mathbb{R}, X_4 \geq c_2, X_5 \in \mathbb{R}\}$. The probability of this event can be written as an integral of the marginal joint density of (X_1, X_2, X_4) which can either be deduced from the joint distribution (2.5) of the whole sequence or found directly, using

the fact that for any $i > j$, $X_i - X_j$ is independent of $X_1, \ldots, X_j$.

For each $k = 1, \ldots, K-1$, define

$$\begin{aligned}
&\psi_k(l_1, u_1, \ldots, l_{k-1}, u_{k-1}, l_k, u_k, c_k; \theta) \\
&\quad = \mathbb{P}(l_1 < S_1 < u_1, \ldots, l_{k-1} < S_{k-1} < u_{k-1}, S_k \geq u_k, \tilde{S}_k \geq c_k; \theta) \\
&\quad + \mathbb{P}(l_1 < S_1 < u_1, \ldots, l_{k-1} < S_{k-1} < u_{k-1}, S_k \leq l_k, \tilde{S}_k \geq c_k; \theta), \qquad (2.6)
\end{aligned}$$

and

$$\begin{aligned}
&\xi_k(l_1, u_1, \ldots, l_{k-1}, u_{k-1}, l_k, u_k, c_k; \theta) \\
&\quad = \mathbb{P}(l_1 < S_1 < u_1, \ldots, l_{k-1} < S_{k-1} < u_{k-1}, S_k \leq l_k, \tilde{S}_k < c_k; \theta) \\
&\quad + \mathbb{P}(l_1 < S_1 < u_1, \ldots, l_{k-1} < S_{k-1} < u_{k-1}, S_k \geq u_k, \tilde{S}_k < c_k; \theta). \qquad (2.7)
\end{aligned}$$

We also define

$$\begin{aligned}
&\psi_K(l_1, u_1, \ldots, l_{K-1}, u_{K-1}, c_K; \theta) \\
&\quad = \mathbb{P}(l_1 < S_1 < u_1, \ldots, l_{K-1} < S_{K-1} < u_{K-1}, \tilde{S}_K \geq c_K; \theta) \qquad (2.8)
\end{aligned}$$

and

$$\begin{aligned}
&\xi_K(l_1, u_1, \ldots, l_{K-1}, u_{K-1}, c_K; \theta) \\
&\quad = \mathbb{P}(l_1 < S_1 < u_1, \ldots, l_{K-1} < S_{K-1} < u_{K-1}, \tilde{S}_K < c_K; \theta). \qquad (2.9)
\end{aligned}$$

Then, a test's power and type I error rate under a given value of θ can be expressed as the sum of the probabilities shown in (2.6) - (2.9), which in turn can be written as integrals of the joint density of $\{S_1, \ldots, S_k, \tilde{S}_k\}$. Since these probabilities cannot be evaluated analytically, in order to compute a test's properties they must be computed by numerical integration.

Referring again to the three-stage test illustrated in Figure 2-3, we now explain how we can exploit the independent increments structure of the sequence of test statistics generated by a delayed response GST to simplify the numerical calculations required to compute probabilities (2.6)-(2.9). Test statistics $\tilde{S}_1$ and S_1 are based on nested datasets and $\tilde{S}_1 - S_1$ is independent of S_1. Similarly, $S_2 - S_1$ is also independent of S_1. For $k = 2, \ldots, K-1$, let $\Delta_k = I_k - I_{k-1}$. Define $\tilde{\Delta}_k = \tilde{I}_k - I_k$, for $k = 1, \ldots, K-1$. Then, conditional on $S_1 = s_1$, $\tilde{S}_1$ and S_2 are distributed according to

$$\tilde{S}_1 | S_1 = s_1 \sim N(s_1 + \theta\tilde{\Delta}_1, \tilde{\Delta}_1),$$

and should we choose to continue to interim analysis $k = 2$,

$$S_2|S_1 = s_1 \sim N(s_1 + \theta\Delta_2, \Delta_2).$$

Following from the independent increments property, $\tilde{S}_2 - S_2$ and $\tilde{S}_3 - S_2$ are independent of both S_2 and S_1. Hence, the conditional densities of $\tilde{S}_2$ and $\tilde{S}_3$ given $S_2 = s_2$ and $S_1 = s_1$ depend only on s_2 and are given by

$$\begin{aligned}\tilde{S}_2|S_1 = s_1, S_2 = s_2 &\sim N(s_2 + \theta\tilde{\Delta}_2, \tilde{\Delta}_2)\\ S_3|S_1 = s_1, S_2 = s_2 &\sim N(s_2 + \theta\Delta_3, \Delta_3).\end{aligned}$$

The arguments presented above in the case $K = 3$ hold for general K; at any interim analysis k, whether we choose to terminate recruitment or not, the increments $\tilde{S}_k - S_k$ and $S_{k+1} - S_k$ are independent of $S_1, \ldots, S_k$. For $k = 1, \ldots, K-1$, let $f_k(\tilde{s}_k|s_k;\theta)$ be the conditional density of $\tilde{S}_k$ given $S_k = s_k$ which is equal to

$$f_k(\tilde{s}_k|s_k;\theta) = \frac{1}{\sqrt{\tilde{\Delta}_k}}\phi\left(\frac{\tilde{s}_k - (s_k + \theta\tilde{\Delta}_k)}{\sqrt{\tilde{\Delta}_k}}\right),$$

where $\phi(x) = \exp(-x^2/2)/\sqrt{2\pi}$ is the density at x of a standard normal variate. Let $f_K(\tilde{s}_K|s_{K-1};\theta)$ be the conditional density of $\tilde{S}_K$ given $S_{K-1} = s_{K-1}$. Similarly, let $g_1(s_1;\theta)$ be the density at s_1 of S_1 and let $g_k(s_k|s_{k-1};\theta)$, for $k = 2, \ldots, K-1$, be the conditional density of S_k at s_k given $S_{k-1} = s_{k-1}$.

Using the independent increments structure, for each $k = 1, \ldots, K-1$, the first term of $\psi_k(l_1, u_1, \ldots, l_{k-1}, u_{k-1}, l_k, u_k, c_k;\theta)$ is

$$\int_{l_1}^{u_1}\cdots\int_{l_{k-1}}^{u_{k-1}}\int_{u_k}^{\infty}\int_{c_k}^{\infty} g_1(s_1;\theta)g_2(s_2|s_1;\theta)\ldots g_k(s_k|s_{k-1};\theta)f_k(\tilde{s}_k|s_k;\theta)ds_1\ldots ds_k d\tilde{s}_k$$

$$\begin{aligned}= \int_{l_1}^{u_1}\cdots\int_{l_{k-1}}^{u_{k-1}}\int_{u_k}^{\infty} & g_1(s_1;\theta)g_2(s_2|s_1;\theta)\ldots g_k(s_k|s_{k-1};\theta)\\ &\times\ \Phi\left(\frac{s_k + \theta\tilde{\Delta}_k - c_k}{\sqrt{\tilde{\Delta}_k}}\right)ds_1\ldots ds_k, \quad (2.10)\end{aligned}$$

where Φ is the cdf of a standard normal variate.

Looking at the integral on the right hand side (RHS) of (2.10), we see that the integrand is a product of terms where each variable of integration appears in only two adjacent factors. Hence, using numerical integration we can approximate the

integral by a k-fold sum which can be evaluated by computing a sequence of two-dimensional sums (for further details see Jennison (1994)). Under this approach, the number of operations required is linear in k. This is an efficient alternative to the MULNOR routine of Schervish (1984) for computing multivariate normal probabilities for which the computational load increases exponentially with k. The same numerical integration routine as outlined above can also be used to compute the second term of $\psi_k(l_1, u_1, \ldots, l_{k-1}, u_{k-1}, l_k, u_k, c_k; \theta)$, which can also be written as a k-fold multiple integral. For each $k = 1, \ldots, K-1$, $\xi_k(l_1, u_1, \ldots, l_{k-1}, u_{k-1}, l_k, c_k; \theta)$ can be expressed as the sum of two k-fold multiple integrals, while $\psi_K(l_1, u_1, \ldots, l_{K-1}, u_{K-1}, c_K; \theta)$ and $\xi_K(l_1, u_1, \ldots, l_{K-1}, u_{K-1}, c_K; \theta)$ can be expressed as $(K-1)$-fold multiple integrals.

One particular variant on the testing scenario outlined in Section 2.2.1 which is of substantial practical interest is the case where short-term measurements known to be correlated with the long term endpoint of interest are also available for each subject. In Section 2.3 below, we fit a joint model for both endpoints and show that the amount of information in the pipeline at an interim analysis is reduced if available responses on this short-term endpoint are incorporated into the analysis. We shall extend to that case the distributional results presented in this section for tests based on measurements on a single long term endpoint. We conclude that the tests that will be derived in subsequent chapters can also be applied when data on a correlated short-term endpoint are available. We finish the next section by illustrating the methodology by means of a simple example.

2.3 Short-term endpoints

2.3.1 Methodology

When the delay in the subject response, Δ_t, is large, it is likely that a short-term measurement which is a good predictor of the eventual response will be available. This may be an early measurement on the clinical endpoint of interest. Alternatively, we may make use of a so-called surrogate endpoint: a biomarker, or physical sign, which is predictive of the clinically meaningful endpoint. For example, in a trial into cancer related bone lesions cited by Todd & Stallard (2005) it was known that early changes in a chemical biomarker were highly predictive of whether or not a subject would eventually experience a skeletal event within 40 weeks, the primary endpoint of clinical interest. While it is assumed that the long term responses of all recruited subjects will be observed before deciding whether or not to claim superiority for the new treatment, making use of short-term measurements may help us to predict how those subjects in the pipeline at an interim analysis will eventually respond.

The delayed response GSTs in the form (2.2) can be still be applied to the data generated by making repeated measurements on each subject in this way. To illustrate this, consider an example similar to that described by Todd & Stallard. Suppose a surrogate endpoint measured early in the study is known to be correlated with the primary endpoint of interest. Responses on different subjects are assumed to be independent. Let $X_{i,1}$ and $X_{i,2}$, $i = 1, \ldots, n_{max}$, denote the response of the ith subject for the surrogate and primary endpoints respectively. Let $T(i) \in \{A, B\}$ be the treatment indicator function such that $T(i) = A$ if subject i receives the experimental treatment and $T(i) = B$ otherwise. Suppose responses on a subject can be modelled as shown below

$$\begin{pmatrix} X_{i,1} \\ X_{i,2} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{T(i),1} \\ \mu_{T(i),2} \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \tau\sigma_1\sigma_2 \\ \tau\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right) \quad \text{for } i = 1, \ldots, n_{max}.$$

We assume at first σ_1^2, σ_2^2 and τ are known. In this setting, τ is the correlation between the short and long term responses. Let $\boldsymbol{\beta} = (\mu_{A,1}, \mu_{B,1}, \mu_{A,2}, \mu_{B,2})^T$ denote our vector of parameters. The efficacy of the new treatment will be measured only by its effect in the clinically relevant endpoint. Hence, $\theta = \mu_{A,2} - \mu_{B,2}$ is the parameter of interest. We wish to test the null hypothesis $H_0 : \theta \leq 0$ against the one-sided alternative $H_1 : \theta > 0$ using a K-stage delayed response GST. At any given interim analysis, not all measurements will have been made on each recruited subject. We refer to those subjects for whom both endpoints are available as "full responders". Those subjects in the pipeline, whose long term response has not been observed, will fall into one of two groups: those whose short-term endpoint is available and those who remain entirely unobserved. Let $\mathbf{X}^{(k)}$ denote the vector of all responses available at interim analysis k. Denote the design and variance-covariance matrices for the data available by $\mathbf{D}^{(k)}$ and $\boldsymbol{\Sigma}^{(k)}$ respectively, so that the data follow the model shown below.

$$\mathbf{X}^{(k)} \sim N(\mathbf{D}^{(k)}\boldsymbol{\beta}, \boldsymbol{\Sigma}^{(k)}).$$

Let $\hat{\boldsymbol{\beta}}_k$ denote the maximum likelihood estimator for the parameter vector $\boldsymbol{\beta}$ based on the data accumulated at interim analysis k. We are only concerned with making inferences on θ. Since we can express $\theta = \mathbf{c}^T\boldsymbol{\beta}$, where $\mathbf{c} = (0, 0, 1, -1)^T$ is our contrast vector, we can extract $\hat{\theta}_k$, the current maximum likelihood estimate of θ using $\hat{\theta}_k = \mathbf{c}^T\hat{\boldsymbol{\beta}}_k$. The information for θ at interim analysis k is given by

$$I_k = \{Var(\mathbf{c}^T\hat{\boldsymbol{\beta}}_k)\}^{-1}. \tag{2.11}$$

The standardised test statistic for testing H_0 is given as usual by $Z_k = \sqrt{I_k}\hat{\theta}_k$. Let $\tilde{I}_k$ and $\tilde{Z}_k$ denote the corresponding quantities at decision analysis $k = 1, \ldots, K$.

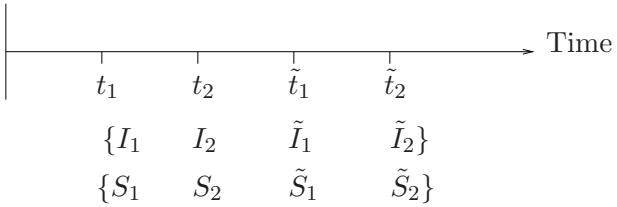


Figure 2-4: An example of the configuration of time and information levels generated by a delayed response GST when measurements on a short-term endpoint are available for each subject.

2.3.2 Distributional results

In this section, we show that the distributional results presented in Section 2.2.2 can be extended to the testing scenario where both short and long term measurements are made on each subject. In this case, the set of data available at an interim analysis is made up by the short and long term responses that have been observed. It is important to recall that at an interim analysis, not all responses will be available for all recruited subjects; for some subjects only short-term measurements will be available, while for others there will be none.

In Section 2.2.2, we assumed only one long-term measurement was made on each subject. In this case, it was clear that the datasets, $\{\mathcal{F}_1, \ldots, \mathcal{F}_{2K-1}\}$, corresponding to the sequence of ordered score statistics, $\{X_1, \ldots, X_{2K-1}\}$, are nested. However, in the case where both short and long term measurements are made, this structure does not necessarily hold. For example, consider the scenario illustrated in Figure 2-4 where $\tilde{I}_1 > I_2$. Suppose recruitment is closed at the first interim analysis. By the time of the decision analysis, the short and long term responses of all those subjects recruited at time $t = t_1$ will have been observed. Contrast this to the case where recruitment is continued at the interim analysis. If the lag in the short-term endpoint is sufficiently short, observations on this endpoint will be available at the next interim analysis for a certain portion of those subjects recruited in the interval $[t_1, t_2)$. It is clear that these observations would not be available at decision analysis 1, as recruitment was closed at time $t = t_1$. Hence, we see that the dataset upon which S_2 is based will not be a subset of that corresponding to $\tilde{S}_1$.

Following these arguments, we see that for the case where short and long term measurements are made on subjects, while properties (i) and (ii) of the canonical distribution (2.5) will continue to hold, we cannot deduce that the standard covariance structure will hold for the whole sequence $\{X_1, \ldots, X_{2K-1}\}$. However, from probabilities (2.6)-(2.9), we see that the boundary calculations for a delayed response GST are dependent only on the marginal joint distribution of the sequences

$\{S_1, \ldots, S_k, \tilde{S}_k\}$, for each $k = 1, \ldots, K-1$, and $\{S_1, \ldots, S_{K-1}, \tilde{S}_K\}$. Hence, we need only consider the joint distribution of each of these sequences. It is clear that, for each k, the data sets from which the statistics in the sequence $\{S_1, \ldots, S_k, \tilde{S}_k\}$ are derived, will be nested. Applying Theorem 1 of Jennison & Turnbull (1997) to the score statistics generated by a sequence of maximum likelihood estimators for a parameter vector in a normal linear model with known variance-covariance matrix, we can deduce that, marginally, this sequence will be multivariate normal with standard covariance in terms of information levels $\{I_1, \ldots, I_k, \tilde{I}_k\}$. Hence, our sequence of score statistics have independent increments and, following the approach taken in Section 2.2.2, we can work directly from this property to show that we get the sequence of conditional probabilities in, say, the RHS of (2.10) in just the right way. We obtain $S_1 \sim g_1(s_1; \theta)$, $S_2|S_1 = s_1 \sim g_2(s_2|s_1; \theta)$ etc, and $\tilde{S}_k|S_k = s_k, \ldots, S_1 = s_1 \sim f_k(\tilde{s}_k|s_k; \theta)$, as required. Hence, our stopping probabilities (2.6)-(2.9) can be written as integrals of the form (2.10).

In the following section, we present a simple example to highlight the gains that can be made if a suitable short-term endpoint can be found.

2.3.3 An example

Suppose we have a new treatment for hypertension which we wish to compare against control. We are interested in investigating whether the treatment is effective at controlling hypertension long-term and so the primary endpoint is change from baseline in mm Hg (millimetres of mercury) after 10 weeks. It is known that change from baseline at two weeks is predictive of this final outcome. Let $X_{i,1}$ and $X_{i,2}$, $i = 1, 2, \ldots$, denote the responses of those subjects allocated to the new treatment at 2 and 10 weeks respectively. Let $Y_{i,1}$ and $Y_{i,2}$, $i = 1, 2, \ldots$, denote the corresponding responses for subjects allocated to control. Suppose

$$\begin{pmatrix} X_{i,1} \\ X_{i,2} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{A,1} \\ \mu_{A,2} \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \tau\sigma_1\sigma_2 \\ \tau\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right),$$

$$\begin{pmatrix} Y_{i,1} \\ Y_{i,2} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{B,1} \\ \mu_{B,2} \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \tau\sigma_1\sigma_2 \\ \tau\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right) \quad \text{for } i = 1, \ldots, n_{max}. \tag{2.12}$$

It is known $\sigma_1^2 = 1$ and $\sigma_2^2 = 2$, and the short and long term measurements are positively correlated with $\tau = 0.7$. We wish to test the hypothesis $H_0 : \theta = \mu_{A,2} - \mu_{B,2} \leq 0$ against the one-sided alternative $H_1 : \theta > 0$ using a two-stage delayed response GST. We set the maximum sample size of the study $n_{max} = 100$ subjects, with equal allocation to each treatment. Recruitment occurs at a constant rate such that it will be completed in

	I_1	$\tilde{I}_1$	$\tilde{I}_2$
No second endpoint	5.0	7.5	12.5
With second endpoint	5.814	7.5	12.5

Table 2.1: Information sequences generated by a two-stage test of $H_0 : \theta \leq 0$ against $H_1 : \theta > 0$ when measurements on a short-term endpoint are and are not available for each subject. Tests are designed with $n_{max} = 100$, $\sigma_2^2 = 2$ and analyses are scheduled according to (2.3).

50 weeks; in the terminology introduced in Section 2.2.1, we are working under $r = 0.2$.

Table 2.1 contrasts the information levels that will be observed at each analysis of our example with those that would be accrued if measurements on a short-term endpoint were not available. At interim analysis $k = 1$, 40 subjects are fully observed. Of the 20 subjects in the pipeline, the short-term responses of 16 are available. Hence, using both the short and long term data, $I_1 = 5.81$, an increase of over 16% on what would be observed if no short-term measurements were made. If recruitment is closed at this first stage, we must wait for all recruited subjects to be fully observed before conducting a decision analysis at information level $\tilde{I}_1 = 7.5$. Since four subjects are completely unobserved at the time of the interim analysis, the limit of I_1 as τ approaches 1 is 7; even if one could impute exactly the missing long term responses of the partially observed subjects from their short-term measurements, 0.5 units of information would still remain in the pipeline. For example, when $\tau = 0.9$, $I_1 = 6.51$, and as τ increases to 0.95 this information level increases to $I_1 = 6.74$. In our example, we have $I_1/\tilde{I}_1 = 0.78$ when the second endpoint is used. The same ratio of I_1 to $\tilde{I}_1$ would be observed if we did not make short-term measurements but the primary endpoint was the change from baseline after 6.3 weeks, in which case $r = 0.13$ instead of $r = 0.2$. Thus, use of a short-term endpoint should help reduce adverse effects of a long delay in observing the primary response. In Chapter 7, we explore in further detail the efficiency gains to be made by making short-term measurements on each subject.

2.4 A road map for the delayed response problem

In this chapter, we have formulated group sequential designs for delayed responses which recognise that the response delay and can be planned ahead of time. In Chapters 3 - 9, we proceed to deal with many facets of the delayed response problem. In Chapter 3, optimal delayed response GSTs are derived which minimise the expected number of subjects recruited on termination under certain scenarios. Comparing these tests against their fixed sample counterparts enables us to quantify the reductions on the fixed sample size that are possible for a delayed response GST. In Chapter

5, we resolve the important issue of how one can analyse the data on termination of a delayed response GST; exact p-values and confidence intervals which attain their nominal coverage rates are derived. The methodology presented in these two chapters is derived under the simplifying assumption that tests are conducted under a fixed sequence of information levels. For practical usefulness, it is important to deal with unpredictable sequences of information however. Our solution, presented in Chapter 6, is to adapt the delayed response test structure using an error spending approach.

The development of error spending methods signals the culmination of our treatment of the delayed response problem in user friendly designs which are flexible enough to be used in practice. These versions can be used in the maximum information monitoring approach of Mehta & Tsiatis (2001), developed in the context of standard GSTs, to cope with delayed responses with unknown variances. Such an approach is vital for our designs to be viable in practice, where nuisance parameters are usually not known exactly. In Chapter 7, we consider the case where measurements on a short and long term endpoint are available for each subject. The accuracy of an information monitoring approach in controlling error rates at their nominal values is assessed via simulation when the variance of each response variable and their correlation is unknown.

In practice, we may not always choose the test minimising expected sample size. Often, achieving a rapid time to a conclusion is just as important. Minimising the expected time to a decision will speed up the drug development process; if the drug is effective we can reduce the time taken for it to reach market, while if it is futile, we can abandon a lost cause and divert resources to other, more promising treatments. Motivated by these considerations, in Chapter 8 we derive tests minimising expected time to a conclusion and also tests optimal for a combination of objectives. We finish our treatment of the delayed response problem in Chapter 9 with a look at deriving tests minimising expected time to a conclusion with survival data. By comparing the performances of these tests with that of standard error spending designs, we are able to make connections between standard group sequential designs efficient for expected sample size and designs which minimise expected time to a conclusion with survival data.

Chapter 3

Deriving optimal delayed response group sequential tests

3.1 Motivation

In Chapter 2, we derived a new group sequential test structure for dealing with delayed responses. However, at present we have not discussed how one might choose the critical values defining this test. In the following section, it is assumed that our primary objective is to choose our test of H_0 so that we minimise the number of subjects required. This scenario would indeed apply if there was a shortage of suitable subjects available to be recruited into the trial; the condition the experimental treatment is intended to treat may be rare or the trial objective may be to confirm the efficacy of the drug in a small subgroup of the main population. As discussed in Chapter 1, economic considerations have also recently assumed a heightened importance, compelling us to seek out testing strategies which make more efficient use of subjects, particularly if the new treatment regimen is expensive to implement. On a more technical note, it was noted in Section 2.1 that the efficiency gains usually quoted for group sequential tests over their fixed sample counterparts are derived in the immediate response setting. Finding optimal delayed response GSTs will allow us to find what reductions on the fixed sample size are possible for a delayed response.

Let N represent the number of subjects recruited on termination of a delayed response GST. Clearly N is a random variable, since the stage at which a group sequential test terminates is data dependent. Our objective is to derive delayed response GSTs for testing $H_0 : \theta \leq 0$ against $H_1 : \theta > 0$ which are optimal with respect to certain criterion while also satisfying given error probability constraints, namely that the test have type I error probability α at $\theta = 0$ and power $1 - \beta$ at $\theta = \delta$. In particular, we

want to find tests minimising objective functions F_i, $i = 1, \ldots, 4$, as defined below

$$F_1 = \mathbb{E}(N; \theta = \delta/2),$$

$$F_2 = 0.5\left\{\mathbb{E}(N; \theta = 0) + \mathbb{E}(N; \theta = \delta)\right\},$$

$$F_3 = 0.5\left\{\mathbb{E}(N; \theta = -\delta/2) + \mathbb{E}(N; \theta = 3\delta/2)\right\},$$

$$F_4 = \int \mathbb{E}(N; \theta)\frac{2}{\delta}\phi\left(\frac{\theta - \delta/2}{\delta/2}\right)\, d\theta,$$

where $\phi(x) = \exp(-x^2/2)/\sqrt{2\pi}$, the standard normal probability density function. Each of the above objective functions has been the subject of previous study. Objective function F_1 represents a worst case scenario where θ is equidistant between the values of the unknown parameter at which the size and power of the test are specified. Minimisation of this objective function when $\alpha = \beta$ is known as the Kiefer-Weiss (Kiefer & Weiss (1957), Weiss (1962)) problem and has been considered in both the fully sequential and group sequential testing paradigms; Lai (1973) and Lorden (1976) find designs in the fully sequential case while Eales & Jennison (1992) and Barber & Jennison (2002) work with GSTs. Objective functions F_2 and F_3 consider average expected sample sizes calculated under increasingly extreme values of θ in the null and alternative parameter spaces. In contrast, F_4 considers the weighted average performance of a test over a continuum of θ values, where the expected sample size is integrated over a $N(\delta/2, (\delta/2)^2)$ distribution for θ. We seek tests which have a good all round performance so that they are efficient by other criteria too. Tests minimising objective functions F_2 and F_4 are of particular interest; Eales & Jennison (1992) find that when $\alpha = \beta$, symmetric GSTs minimising F_2 and F_4 perform close to optimal with regard to other objective functions.

The problem of finding optimal tests satisfying given error probability constraints is a constrained minimization problem. We adopt the approach taken by Eales & Jennison (1992) and reformulate the problem using Bayesian sequential decision theory, where the required optimal frequentist test is found as the solution to an unconstrained Bayes problem. Banerjee & Tsiatis (2006) point out that this is equivalent to finding the solution using Lagrangian multipliers. Our approach is explained in more detail in the following section.

3.2 Formulation of the optimisation problem

To recapitulate, our objective is to find the K stage delayed response GST for testing $H_0 : \theta \leq 0$ minimising objective function F_i, $i = 1, \ldots, 4$, while also attaining type I error probability α at $\theta = 0$ and power $1 - \beta$ at $\theta = \delta$. We illustrate the methodology used to find such tests by first considering the minimisation of F_1 in some detail. In the following discussion, we fix parameters K, α, β, δ, and also our sample size inflation factor R, where $n_{max} = Rn_{fix}$. We also fix the delay parameter r rather than work directly in terms of Δ_t.

When trying to find optimal tests, it is useful to think of our delayed response GST as a bounded sequential decision procedure. Each analysis can be regarded as a decision node where we must choose between several courses of action. The decision problem is bounded, since in the absence of early stopping, a decision must be made once all n_{max} subjects have been accrued and their responses observed. Contrast this to an "open" procedure such as the SPRT (Wald, 1947); here the decision problem is unbounded because sampling can continue indefinitely until the test statistic sample path crosses a stopping boundary. When thinking of our test as a decision procedure, it seems natural to formalise our decision making process so that in light of our overall objective of minimising F_1, one can choose the best possible course of action at each decision node. We do this by using three essential ingredients to reformulate our problem as a Bayes sequential decision problem; namely, we use a prior distribution, π, for the unknown parameter, a sampling cost function, $c(\theta)$, and a decision loss function $L(A, \theta)$, which is a function of the action, A, taken and the true θ value.

In order to minimise F_1, we place a uniform discrete three point prior on θ such that $\pi(0) = \pi(\delta/2) = \pi(\delta) = 1/3$. When defining the loss function we note that our one-sided test of H_0 is asymmetric in the sense that α is not constrained to equal β. Let A_i denote the action of accepting hypothesis i, $i = 0, 1$. Set $L(A_1, 0) = d_1$, $L(A_0, \delta) = d_0$ and all other $L(A, \theta) = 0$. We can interpret d_1 and d_0 as the losses incurred by making a type I error under $\theta = 0$ and a type II error under $\theta = \delta$ respectively, where both decision costs must be positive. Since our objective is to minimise a function of N, our sampling cost function is defined such that we are charged $c(\theta)$ units per subject recruited. For objective function F_1, we set $c(\delta/2) = 1$ and $c(\theta) = 0$ otherwise. Define $\Theta = \{0, \delta/2, \delta\}$ to be the parameter space for θ, T to represent the stage at which the test terminates, and write the decision made at stage k in the form $\lambda(z_1, \ldots, z_k, \tilde{z}_k)$ to stress that the decisions made are functions of the data. Also define $\{T = k\}$ to be the set of sample paths such that the GST terminates at stage k, $k = 1, \ldots, K$. The total

expected cost of the trial can be written as

$$\sum_{\Theta}\sum_{k=1}^{K}\int_{\{T=k\}} L(\lambda(z_1,\ldots,z_k,\tilde{z}_k),\theta)f_k(z_1,\ldots,z_k,\tilde{z}_k|\theta)\pi(\theta)\,dz_1,\ldots,dz_k d\tilde{z}_k + \mathbb{E}[\,c(\theta)N\,] \tag{3.1}$$

where $f_k(z_1,\ldots,z_k,\tilde{z}_k|\theta)$ is the $(k+1)$th dimensional joint density of $Z_1,\ldots,Z_k,\tilde{Z}_k$ at $z_1,\ldots,z_k,\tilde{z}_k$ given the true effect size is θ. A test minimising the total expected cost is said to be a Bayes test for this problem. Noting that

$$\mathbb{E}[\,c(\theta)N\,] = \mathbb{E}_\theta[\,c(\theta)\mathbb{E}(N|\theta)\,],$$

where $\mathbb{E}_\theta[\,g(\theta,N)\,]$ denotes taking expectations of the function g with respect to the distribution of θ, simplifying (3.1) we find the total expected cost of the test is

$$\pi(0)d_1\mathbb{P}(\text{Accept } H_1|\theta=0)+\pi(\delta)d_0\mathbb{P}(\text{Accept } H_0|\theta=\delta)+\pi(\delta/2)\mathbb{E}(N|\theta=\delta/2). \tag{3.2}$$

For a given pair of costs (d_0,d_1), let $\alpha^\star$ denote the type I error rate at $\theta=0$ and $\beta^\star$ the type II error rate at $\theta=\delta$ for the test minimising (3.2). Following the usual Lagrangian argument, a test minimising (3.2) must also minimise F_1 among tests with the same error rates $\alpha^\star$ and $\beta^\star$. However, this test may not necessarily have the required error probabilities. Since the error probabilities of a Bayes test minimising (3.2) depend on our decision costs d_0 and d_1, all that remains is to search for the pair of costs $(d_0^\star,d_1^\star)$ defining a Bayes problem with a solution which has error rates $\alpha^\star=\alpha$ and $\beta^\star=\beta$. This Bayes test minimises F_1 amongst all tests with error rates α and β and so is a solution to our original problem. Define $f_1(a,b)$ and $f_2(a,b)$ respectively to be the type I error rate at $\theta=0$ and type II error rate at $\theta=\delta$ of the Bayes test corresponding to the pair of decision costs $(d_0=\exp(a),d_1=\exp(b))$. The decision costs $d_0^\star$ and $d_1^\star$ are found as the solutions to the simultaneous equations

$$\begin{aligned}\{f_1(\log(d_0),\log(d_1))-\alpha\}^2&=0\\ \{f_2(\log(d_0),\log(d_1))-\beta\}^2&=0.\end{aligned} \tag{3.3}$$

We regard f_1 and f_2 as functions of log decision costs for computational convenience because it means that we can find $(d_0^\star,d_1^\star)$ using an unconstrained search over $\mathbb{R}^2$.

As a brief aside, note that it is not intended that the prior for θ reflect investigators initial beliefs. Indeed, the test found minimising (3.2) under prior π and decision costs $(d_{0,1},d_{1,1})$ can also be found as the solution to the Bayes problem defined under any other choice of three-point prior, π', placing weights on $\theta=0$, $\delta/2$ and $\theta=\delta$, if we also

change our decision costs in tandem to

$$d_0 = \frac{\pi'(\delta/2)d_{0,1}}{\pi'(\delta)} \qquad d_1 = \frac{\pi'(\delta/2)d_{1,1}}{\pi'(0)}.$$

For convenience, we shall use $\pi(0) = \pi(\delta/2) = \pi(\delta) = 1/3$ from now on.

All that remains for us to show is how one can find a Bayes test minimising the total expected cost of a trial under a given pair of decision costs (d_0, d_1). In the next section, we explain how this is possible using the technique of backwards induction.

3.3 Finding optimal tests

3.3.1 Backwards induction

As noted in Section 3.2, at each analysis of the GST we have to choose between different courses of action. Each action, or decision, will be associated with an expected additional cost incurred if that action is selected. The optimal decision is that associated with the lowest expected additional cost. It is known that the test minimising the total expected cost of the trial given in (3.1) will be the decision procedure where the optimal decision is made at each stage (DeGroot, 1979). If we are working forwards in time it is not always possible to determine the expected additional cost associated with each decision. For example, it is not possible to calculate whether it is optimal to collect more information at an interim analysis until it is known how this information will be subsequently used. In the context of our delayed response GSTs, this can be assessed since the trial has a finite future, i.e. in the absence of early stopping it must terminate at the end of stage K. The above logic implies that to find the optimal test one must apply a technique called backwards induction; beginning at the final stage of the test we must work backwards, finding the optimal action at each stage using knowledge of the optimal future design the trial will follow if it is decided to continue. Below, we explain in more detail how backwards induction can be used to find the boundary constants $\{l_1, u_1, c_1, \ldots, l_K, u_K, c_K\}$ defining an optimal test for a general unconstrained Bayes problem.

Note that for each k, $\tilde{Z}_k$ and Z_k are the sufficient statistics for θ at decision and interim analysis k respectively. Denote the posterior for θ at interim analysis k given $Z_k = z_k$ by $\pi^{(k)}(\theta|z_k)$ and denote the posterior at decision analysis k given $\tilde{Z}_k = \tilde{z}_k$ by $\tilde{\pi}^{(k)}(\theta|\tilde{z}_k)$.

Recall that if recruitment is terminated at interim analysis $k = 1, \ldots, K-1$, we must

wait to observe the pipeline subjects before making a final hypothesis decision at decision analysis k. At this analysis we can either accept H_0 or H_1. Suppose we observe $\tilde{Z}_k = \tilde{z}_k$. The expected loss associated with decision A_i, $i = 0, 1$, is defined to be the expectation of $L(A_i, \theta)$ with respect to the current posterior distribution of θ. The expected loss associated with the optimal decision is given by

$$\eta^{(k)}(\tilde{z}_k) = \min\{d_1\tilde{\pi}^{(k)}(0|\tilde{z}_k), d_0\tilde{\pi}^{(k)}(\delta|\tilde{z}_k)\}.$$

The critical value at decision analysis k, denoted c_k, is found by solving

$$d_1\pi^{(k)}(0|c_k) = d_0\pi^{(k)}(\delta|c_k). \tag{3.4}$$

We see our decision constant c_k is the value of $\tilde{z}_k$ at which the expected loss functions associated with the actions reject H_0 and accept H_0 intersect. Equation (3.4) can be solved analytically for c_k to obtain

$$c_k = \frac{1}{\sqrt{\tilde{I}_k}}\left(\frac{\log(d_1/d_0)}{\delta} + \frac{\delta\tilde{I}_k}{2}\right) \quad \text{for } k = 1, \ldots, K. \tag{3.5}$$

At interim analysis $k = 1, \ldots, K-1$, the optimal testing procedure can continue in one of two ways, namely

1. Terminate recruitment, wait for all pipeline subjects to respond and make the optimal decision given the final dataset. We refer to this as action (1).
2. Continue recruitment until interim analysis $k+1$ and proceed optimally thereafter. We refer to this as action (2).

As an aside, note that when setting up our delayed response group sequential test structure, we imposed upon ourselves the constraint that if at an interim analysis recruitment is terminated, all pipeline subjects must be observed before making a final decision. However, because our optimality criteria F_i, $i = 1, \ldots, 4$, are functions of the expected number of subjects recruited, the sampling cost function in the corresponding Bayes problem only charges us per subject recruited. Therefore, when finding the solution to this Bayes problem, even if we allow ourselves the freedom of not having to wait for the pipeline data before making a final decision, it will never be optimal to take advantage of this freedom; we have already "paid" to recruit the pipeline subjects into the trial and the extra information we accrue by waiting to observe them is obtained for "free".

Suppose we are at interim analysis k and observe $Z_k = z_k$. If action (1) is selected, recruitment will be terminated and so no further sampling costs will be incurred. From the discussion above, we know that at the decision analysis, if we observe $\tilde{Z}_k = \tilde{z}_k$

the posterior expected loss associated with the optimal decision is $\eta^{(k)}(\tilde{z}_k)$. However, since $\tilde{Z}_k$ is random, we must take expectations of this quantity with respect to the conditional distribution $\tilde{Z}_k|Z_k = z_k$, so that we only take into consideration those outcomes which are plausible given the current data. Therefore, the total expected additional cost associated with action (1) is given by

$$\begin{aligned}\rho^{(k)}(z_k) = d_1\mathbb{P}(\theta = 0|z_k)\mathbb{P}(\text{Decide Reject } H_0|z_k, \theta = 0)\\ + d_0\mathbb{P}(\theta = \delta|z_k)\mathbb{P}(\text{Decide Accept } H_0|z_k, \theta = \delta),\end{aligned}$$

which can be written as

$$\begin{aligned}\rho^{(k)}(z_k) = d_1\pi^{(k)}(0|z_k)\mathbb{P}(\tilde{Z}_k \geq c_k|z_k, \theta = 0)\\ + d_0\pi^{(k)}(\delta|z_k)\mathbb{P}(\tilde{Z}_k < c_k|z_k, \theta = \delta).\end{aligned} \tag{3.6}$$

The conditional probabilities in (3.6) cannot be evaluated analytically, but can be computed numerically using a standard library routine to calculate probabilities for the standard normal distribution.

We could also choose action (2) at an interim analysis. At interim analysis $k = 1, \ldots, K-2$, suppose after observing $Z_k = z_k$ we continue to stage $k+1$ and act optimally thereafter. Therefore, at interim analysis $k+1$, upon observing $Z_{k+1} = z_{k+1}$, we choose the optimal action associated with the lowest expected additional loss. Clearly Z_{k+1} is random, so we take expectations of this cost with respect to the conditional distribution $Z_{k+1}|Z_k = z_k$. Let $f_{k+1}(z_{k+1}|z_k)$ be the conditional density of Z_{k+1} at z_{k+1} given $Z_k = z_k$ which is a mixture of normals based on the posterior density $\pi^{(k)}$. Note that we incur a sampling cost of one unit per subject recruited only if the true value of θ is $\delta/2$. Using this information, the total expected additional cost associated with action (2) at interim analysis k, for $k = 1, \ldots, K-2$, is given by

$$\begin{aligned}\beta^{(k)}(z_k) = (\tilde{n}_{k+1} - \tilde{n}_k)\pi^{(k)}(\delta/2|z_k)\\ + \int_{z_{k+1}} \min\{\beta^{(k+1)}(z_{k+1}), \rho^{(k+1)}(z_{k+1})\} f_{k+1}(z_{k+1}|z_k)\, dz_{k+1}.\end{aligned} \tag{3.7}$$

At interim analysis $k = K-1$, if action (2) is taken, recruitment is completed and the trial must progress to decision analysis K, at which point a final hypothesis decision will be made. Define $g_K(\tilde{z}_K|z_{K-1})$ to be the conditional density of $\tilde{Z}_K$ at $\tilde{z}_K$ given $Z_{K-1} = z_{K-1}$, a mixture of normal densities based on the posterior $\pi^{(K-1)}$. Then,

given $Z_{K-1} = z_{K-1}$, the total expected additional cost associated with action (2) is

$$\beta^{(K-1)}(z_{K-1}) = (n_{max} - \tilde{n}_{K-1})\pi^{(K-1)}(\delta/2|z_{K-1}) + \int_{\tilde{z}_K} \eta^{(K)}(\tilde{z}_K) g_K(\tilde{z}_K|z_{K-1})\, d\tilde{z}_K. \quad (3.8)$$

Noting that we can write

$$\int_{\tilde{z}_K} \eta^{(K)}(\tilde{z}_K) g_K(\tilde{z}_K|z_{K-1})\, d\tilde{z}_K = d_1\pi^{(K-1)}(0|z_{K-1})\mathbb{P}(\tilde{Z}_K \geq c_K|z_{K-1}, \theta = 0) + d_0\pi^{(K-1)}(\delta|z_{K-1})\mathbb{P}(\tilde{Z}_K < c_K|z_{K-1}, \theta = \delta),$$

we see that $\beta^{(K-1)}(z_{K-1})$ is a linear function of probabilities for a normal random variable, which can be computed numerically using a standard library routine.

3.3.2 Uniqueness of the Bayes test

So far, we have presented methods for finding a solution to a Bayes problem, but as of yet we have not said whether it is the only solution. We know that the Bayes test will not be unique everywhere: there is more than one optimal decision at the point at which the cost curves intersect. However, we now prove the following theorem:

Theorem 1. *The Bayes problem defined by the pair of costs* (d_0, d_1) *will have a unique solution up to sets of Lebesgue measure zero.*

Proof: First suppose the Bayes problem does not have a unique solution. Then, at some stage k we must either have $\rho^{(k)}(z_k) = \beta^{(k)}(z_k)$ on an interval of values of z_k and/or $\tilde{\pi}^{(k)}(0|\tilde{z}_k) = \tilde{\pi}^{(k)}(\delta|\tilde{z}_k)$ on an interval of values of $\tilde{z}_k$. Note that $\rho^{(k)}(z_k)$ and $\beta^{(k)}(z_k)$ are analytic functions of z_k, and $\tilde{\pi}^{(k)}(\theta|\tilde{z}_k)$ is an analytic function of $\tilde{z}_k$. Extending to our case the arguments of Brown et al. (1980, Theorem 3.3), who use properties of analytic functions proved by Farrell (1968, Lemma 4.2), it follows that the interval on which $\rho^{(k)}(z_k) = \beta^{(k)}(z_k)$ must be either a set of measure zero or the whole of $\mathbb{R}$. The same argument applies to the posterior distribution $\tilde{\pi}^{(k)}$ at decision analysis k. Since we have been able to find solutions of the form (2.2) to all the Bayes problems we have considered, we conclude that the risk functions are not equal everywhere. Hence, we claim uniqueness of the Bayes tests in all the examples we have considered up to sets of Lebesgue measure zero. □

Recall from Section 3.2 that the test minimising F_1 with error rates $\alpha^\star = \alpha$ and $\beta^\star = \beta$ is found by searching for the decision costs $(d_0^\star, d_1^\star)$ defining a Bayes problem whose unique solution satisfies the error rate constraints. This test minimises F_1 in the class of tests with the same error probabilities and we claim it must be the unique solution

(up to sets of measure zero) to our original frequentist problem. To see why this should be, suppose there is a different pair of costs (d_0', d_1') defining a Bayes problem whose solution also has error rates $\alpha^\star = \alpha$ and $\beta^\star = \beta$ and achieves the minimum of F_1. Then this second test can only differ from the solution to the Bayes problem with costs $(d_0^\star, d_1^\star)$ by actions on a set of measure zero by the uniqueness of this first test.

In the next section, we describe in more detail the algorithm used to find optimal frequentist tests computationally.

3.3.3 Implementation of backwards induction

Recall that the decision constants c_k, $k = 1, \ldots, K$, can be found analytically, according to equation (3.5). However, to find the critical values (l_k, u_k), $k = 1, \ldots, K-1$, we must implement a dynamic linear programming algorithm. Starting at stage $K-1$, we set up a grid of values for z_{K-1} to which we can apply a numerical integration routine. Following Jennison (1994), we create a grid of $6\omega - 1$ points, $\{z_j, j = 1, \ldots, 6\omega - 1\}$, where ω is the parameter controlling how refined this mesh should be. Jennison & Turnbull (2000, Chapter 19) suggest using $\omega = 16$: they find computed probabilities for standard GSTs to be within 10^{-6} of their true values, although they recommend higher values of ω when there are small increments of information between consecutive analyses. The grid is chosen to be efficient for integrating the marginal density of Z_{K-1} over the whole real line. Two thirds of grid points are concentrated within ± 3 standard deviations of the mean, with logarithmic spacing of grid points thereafter to reflect the fact that the normal density decays quickly in the tails. Simpson rule weights associated with each grid point z_j, $j = 1, \ldots, 6\omega - 1$, are calculated and stored in another mesh. Chandler & Graham (1988) prove that when integrating a normal density under this choice of grid, Simpson's rule achieves $\mathcal{O}(n^{-4})$ convergence as $n = 6\omega - 1$, the total number of grid points, increases; computed integrals should converge to their true values by one decimal place each time ω is doubled. Each objective function F_i, $i = 1, \ldots, 4$, corresponds to a different prior for θ, leading to a different marginal distribution for each Z_k and hence a different grid of points.

Recall that when finding the Bayes test, we need only consider procedures whose stopping rules specify Bayes decisions at each stage. We see that the critical values defining the optimal test must mark the boundaries of intervals corresponding to different optimal actions. Hence, l_{K-1} and u_{K-1} are found as the solutions to

$$\rho^{(K-1)}(z_{K-1}) - \beta^{(K-1)}(z_{K-1}) = 0.$$

We evaluate $\rho^{(K-1)}(z_{K-1}) - \beta^{(K-1)}(z_{K-1})$ for the grid of points for z_{K-1}. Looking

for changes in sign in this expression between successive grid points, we can identify the approximate location of any roots. There should be only two pairs of grid points for which a change of sign occurs if optimal actions occupy intervals in a standard order, i.e. action (1) is optimal for all $z_k \geq u_k$ and $z_k \leq l_k$, otherwise the next group of subjects is recruited. A check is incorporated into our routines to ensure that we are alerted should this condition fail to hold although without exception this has not occurred in any of the examples we have considered. Once a pair a grid points has been identified, a bisection search is implemented under a certain tolerance, tol, such that we find the interval $[a, b]$ in which our critical value must lie, where $b - a < tol$. If tol is suitably small, it is reasonable to assume that $\rho^{(K-1)}(z_{K-1}) - \beta^{(K-1)}(z_{K-1})$ will be approximately linear over this interval and hence the critical value can be found using linear interpolation.

It is clear that our computed critical values will be subject to certain degree of error. If we wish to find them to a given degree of accuracy, using a bisection search followed by linear interpolation is computationally more efficient than using only a bisection search with a smaller tolerance. It also means that our computed critical values, and hence computed error rates, will be a continuous function of the decision costs d_1 and d_0. This is important in ensuring that our numerical search to find $(d_0^\star, d_1^\star)$ defining the Bayes procedure satisfying the frequentist error probability constraints converges. Problems will emerge if, on termination of the bisection search, we set the computed critical value equal to either the upper or lower endpoint of the interval $[a, b]$. Small changes in our decision costs will result in small changes in our critical values. Hence, when ϵ is small, bisection searches conducted under (d_0, d_1) and $(d_0 + \epsilon, d_1 + \epsilon)$ will terminate with the same interval $[a, b]$ and the computed critical values in each case will be set equal to a. However, if ϵ is sufficiently large, a discontinuity occurs in the computed critical value; the bisection search terminates with a different interval $[a', b']$ and the computed critical value jumps to a'. This jump leads to a discontinuity in the computed error rates which in turn can be a problem when trying to trying to search for the pair of decision costs $(d_0^\star, d_1^\star)$ which define a Bayes problem with a solution with error rates $\alpha^\star = \alpha$ and $\beta^\star = \beta$. The numerical method for solving the pair of simultaneous equations (3.3) is by minimisation of

$$\{f_1(\log(d_0), \log(d_1)) - \alpha\}^2 + \{f_2(\log(d_0), \log(d_1)) - \beta\}^2. \tag{3.9}$$

The efficient routines for doing this use estimated derivatives of (3.9); discontinuities in the numerical version of this sum can make these routines fail.

Once l_{K-1} and u_{K-1} have been found, these points are then added to our original grid. For each grid point z_j, we compute $\min\{\beta^{(K-1)}(z_j), \rho^{(K-1)}(z_j)\}$ and store the results

in a separate mesh. This is then used when computing $\beta^{(K-2)}(z_{K-2})$, as shown in (3.7), numerically using Simpson's rule in the next loop of the algorithm. Adding the points l_{K-1} and u_{K-1} to the original grid is required since there is a discontinuity in the first derivative of $\min\{\beta^{(K-1)}(z_{K-1}), \rho^{(K-1)}(z_{K-1})\}$ at $z_{K-1} = l_{K-1}$ and $z_{K-1} = u_{K-1}$. Simpson's rule achieves $\mathcal{O}(n^{-4})$ convergence only if the function multiplying the normal density in the integrand is smooth and bounded. Inserting grid points at these discontinuities ensures that the aimed for rate of convergence holds. The algorithm described above is repeated $(K-2)$ times until all required critical values have been found. The meshes and numerical integration techniques implemented in the programs used to derive the results in this thesis are presented in more depth in Jennison (1994) and Jennison & Turnbull (2000, Chapter 19).

3.4 Other objective functions

We derive delayed response group sequential tests minimising objective functions F_i, $i = 2, 3, 4$, using the same methodology as described above for function F_1, with adjusted priors and sampling cost functions. The loss function for a wrong decision is not altered. For example, when minimising F_2 we set $\pi(0) = \pi(\delta) = 1/2$ and $c(0) = c(\delta) = 1$ and $c(\theta) = 0$ otherwise. Substituting this into (3.1), we see that in this instance the total expected cost of the trial is

$$\frac{1}{2}\{d_1\mathbb{P}(\text{Accept } H_1|\theta = 0) + d_0\mathbb{P}(\text{Accept } H_0|\theta = \delta)\} + F_2.$$

Similarly, for F_3 we set $\pi(-\delta/2) = \pi(0) = \pi(\delta) = \pi(3\delta/2) = 1/4$. We also define $c(-\delta/2) = c(3\delta/2) = 1$, and $c(\theta) = 0$ otherwise. Under these settings, the total expected cost of the trial is

$$\frac{1}{4}\{d_1\mathbb{P}(\text{Accept } H_1|\theta = 0) + d_0\mathbb{P}(\text{Accept } H_0|\theta = \delta)\} + \frac{1}{2}F_3.$$

Finally, for objective function F_4, let $\pi(0) = \pi(\delta) = 1/3$ and set a prior probability of $1/3$ on the scenario that $\theta \sim N(\delta/2, (\delta/2)^2)$. The sampling cost function is set to be $c(\delta) = c(0) = 0$, and $c(\theta) = 1$ otherwise. In this case, the total expected cost of the trial is

$$\frac{1}{3}\{d_1\mathbb{P}(\text{Accept } H_1|\theta = 0) + d_0\mathbb{P}(\text{Accept } H_0|\theta = \delta) + F_4\}.$$

As is the case with F_1, for each of the objective functions F_i, $i = 2, 3, 4$, the set of recursive relations for finding the critical values by backwards induction can be derived following the same lines of reasoning explained in Section 3.3.1.

So far in this thesis, we have formulated a new test structure for delayed responses and

presented methodology for finding optimal versions of these tests. Before we move on to evaluating these designs in some detail, it is perhaps worth pausing to explore in more detail some of the facets of these new group sequential designs. We do this in the next section by means of a simple example.

3.5 An example

Consider a trial designed to compare two treatments A and B. Measurements on the endpoint of direct clinical interest are made 4 weeks after commencement of treatment, i.e., $\Delta_t = 4$. Let $X_{A,i} \sim N(\mu_A, \sigma^2)$ and $X_{B,i} \sim N(\mu_B, \sigma^2)$, $i = 1, 2, \ldots$, represent the responses of those subjects allocated to treatments A and B, respectively. All responses are independent with known variance $\sigma^2 = 1$. Define $\theta = \mu_A - \mu_B$. We wish to design a three-stage GST of $H_0 : \theta \leq 0$ against $H_1 : \theta > 0$ with type I error probability $\alpha = 0.05$ and power $1 - \beta = 0.9$ at $\theta = 1$. Patient entry is to be divided equally between treatments A and B. The fixed sample test for this problem requires information

$$I_{fix} = \{\Phi^{-1}(1 - 0.05) + \Phi^{-1}(1 - 0.9)\}^2 = 8.564.$$

We set the GST's maximum information level $I_{max} = 1.15 I_{fix}$ which equates to a maximum total sample size over both treatments of $n_{max} = 39.4$ subjects. Of course in practice, n_{max} must be integer; rounding n_{max} up to the nearest multiple of 2, we obtain $n_{max} = 40$ and $I_{max} = 10$. It is anticipated that accrual will proceed at a rate of $c = 2$ subjects a week, with randomisation balanced in blocks of two. Hence, recruitment will be completed after $t_{max} = 20$ weeks and $r = \Delta_t / t_{max} = 0.2$. No additional delay is anticipated in processing data ready for an analysis. Eight subjects will be in the pipeline at an interim analysis. We plan our test of H_0 to be of the form (2.2) in order to provide a proper treatment of the overrun data that will accumulate should termination be triggered.

The scheduling of the interim analyses follows (2.3), generating the information sequence $\{I_1 = 3, \tilde{I}_1 = 5, I_2 = 5.5, \tilde{I}_2 = 7.5, \tilde{I}_3 = 10\}$. Figure 3-1(a) shows the boundaries of the optimal test of $H_0 : \theta \leq 0$ against $H_1 : \theta > 0$ minimising F_2 for our problem. Define $\mathcal{C}_k$ to be the continuation region at the kth interim analysis. Then, for our optimal test, $\mathcal{C}_1 = (-0.06, 2.15)$ and $\mathcal{C}_2 = (0.58, 2.04)$. The test begins with the first interim analysis, conducted once 12 responses have been observed. If $Z_1 \notin \mathcal{C}_1$, recruitment is halted: we wait to follow-up the 8 pipeline subjects before rejecting H_0 if $\tilde{Z}_1 \geq 1.35$, and accepting it otherwise. Alternatively, if $Z_1 \in \mathcal{C}_1$, the optimal test stipulates that sampling continue, and 10 additional responses are observed by the next, and final, interim analysis. At this analysis, recruitment is closed early if $Z_2 \notin \mathcal{C}_2$; once the pipeline information becomes available, H_0 is rejected if $\tilde{Z}_k \geq 1.56$,

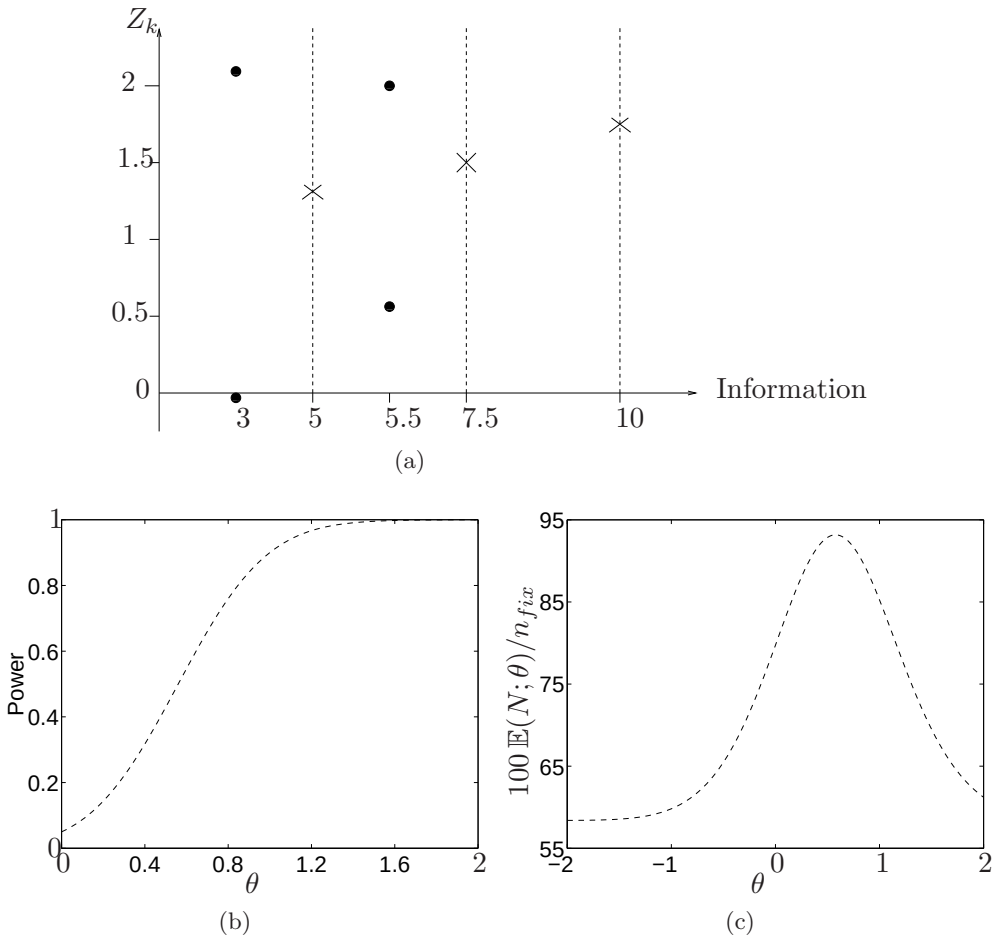


Figure 3-1: Example: (a) Three-stage test of $H_0 : \theta \leq 0$ against $H_1 : \theta > 0$ minimising F_2 for the information sequence $\{I_1 = 3, \tilde{I}_1 = 5, I_2 = 5.5, \tilde{I}_2 = 7.5, \tilde{I}_3 = 10\}$ with $\alpha = 0.05$ and $\beta = 0.1$. (b) Power curve of the optimal test minimising F_2. (c) Curve of the the optimal test's expected sample size under effect size θ expressed as a percentage of the corresponding fixed sample size.

otherwise it is accepted. However, if the test statistic sample path does not exit the continuation region at the second interim analysis, recruitment must remain open until its completion; once all $n_{max} = 40$ responses are available, we reject H_0 if $\tilde{Z}_3 \geq 1.74$, else we accept it. Figure 3-1(c) shows that while our GST requires a larger maximum sample size to match the power of the fixed sample test at $\theta = \delta$, by testing group sequentially we can make large savings in expected sample size, particularly under null values of θ or $\theta \gg \delta$. The expected sample size curve is symmetric about its maximum at around $\theta = 0.56$ when $\mathbb{E}(N;\theta) = 93\%$ of n_{fix}.

There are perhaps certain facets to our new designs requiring further comment. One might ask questions such as "do you ever exit the upper boundary at an interim analysis but then finally accept H_0?" and "are the pipeline observations actually used?". For

our example, for which $r = 0.2$,

$$\mathbb{P}(Z_1 \geq 2.15, \tilde{Z}_1 < 1.35; \theta = \delta) = 0.00270$$
$$\mathbb{P}(-0.063 < Z_1 < 2.15, Z_2 \geq 2.04, \tilde{Z}_2 < 1.56; \theta = \delta) = 0.00264,$$

which are reasonably small values. The conditional probability under $\theta = \delta$ that we reject H_0 given that termination is triggered by observing $Z_1 = 2.15$ at the first interim analysis is 0.972. Under $\theta = 0$, this conditional probability decreases to 0.691. In Chapter 4, we extend these conclusions and find that when r is small, the pipeline observations are little used by optimal tests; in effect we are deciding whether to reject or accept H_0 at the interim analysis. However, it is still important to have a formal framework for dealing with the overrun data, and this is provided by our designs. For larger values of r, the pipeline data do assume a more important role in our final decision of whether to reject or accept H_0; we shall show that in these instances, our new designs really are better than just using a standard GST.

Let $z_\gamma = \Phi^{-1}(1-\gamma)$. Looking at the decision constants for an optimal one-sided test with symmetric error rates $\alpha = \beta$ given by $c_k = (\delta/2)\sqrt{\tilde{I}_k}$, for $k = 1, \ldots, K$, we notice that in the early stages of the test, c_k may well fall below $z_{0.05}$. It comes as little surprise then that in our example with $\alpha \neq \beta$, our optimal test has c_1 and $c_2 < z_{0.05}$. Then, in the event of early stopping, it is possible that the GST rejects H_0 with $\tilde{Z}_k < z_\alpha$. Regulators may be wary of approving a drug if this is the case. One solution would be to impose a constraint $c_k \geq z_\alpha$, although we have not done this in our work. In our optimal tests, for small values of r, given that $Z_k \notin \mathcal{C}_k$, the probability that $\tilde{Z}_k$ takes a value in the neighbourhood of c_k is close to zero. Hence, tests optimised with the additional constraint of $c_k \geq z_\alpha$ would be very similar to our optimal tests and have very similar properties. Therefore, results for the optimal delayed response designs derived in this chapter are still of considerable practical interest and they are presented in Chapter 4.

Chapter 4

Properties of optimal delayed response tests

4.1 Introduction

The efficiency results usually quoted for GSTs are calculated assuming response is immediate. For example, Barber & Jennison (2002) evaluate optimal standard designs with asymmetric error rates when there is no delay in response. They find that savings of around 30% on the fixed sample test are possible with as few as 5 analyses. Table 4.1 lists the minima of F_4 for $r = 0$ when $\alpha = 0.05$ and $\beta = 0.1$. Results are indexed by R, the sample size inflation factor, and K, the maximum number of stages in the GST permitted. For fixed K, we see that the minimum of F_4 decreases and then increases with R. This trend is also replicated for objective functions $F_1, \ldots, F_3$; see too the results of Eales & Jennison (1992), who evaluate optimal standard designs with symmetric error rates assuming response is immediate. For optimal two-stage designs designed under $\alpha = 0.05$ and $\beta = 0.1$, the turning point of the minima occurs at approximately $R = 1.15$ for all objective functions. Table 4.1 shows that even for $K > 2$, we are not far off the global minimum of F_4 if we set $R = 1.15$.

In this chapter, we evaluate the optimal versions of our delayed response group sequential designs which were derived in Chapter 3 to find the benefits for early stopping that are possible when there is a delay in response. Using the rationale given above, all the results presented are derived under $R = 1.15$.

	R					
K	1.1	1.15	1.2	1.3	1.4	1.5
2	77.7	77.5	78.0	79.8	82.2	85.0
3	72.0	71.0	70.6	70.7	71.5	72.7
5	68.1	66.7	65.9	65.3	65.2	65.5
10	65.1	63.6	62.7	61.7	61.3	61.2

Table 4.1: Minima of F_4 when there is no delay in response expressed as a percentage of the corresponding fixed sample size. Tests are derived under $\alpha = 0.05$ and $\beta = 0.1$.

4.2 Properties of optimal delayed response tests

Using the methodology presented in Chapter 3, we have found and evaluated optimal delayed response GSTs for various F_i, K and r. Tables 4.2 to 4.5 list the minima of the objective functions expressed as a percentage of the fixed sample size. Attention is restricted to $2 \leq K \leq 20$, although we have developed software which is capable of finding optimal tests with K as large as 200. As discussed in the preamble to this chapter, all tables are produced under $R = 1.15$ and with analyses scheduled at information levels according to (2.3). The column $r = 0$ is included for reference and corresponds to the case when response is immediate. Provided r is equal to the given value, the results presented in Tables 4.2 to 4.5 are invariant to changes in δ, the alternative at which power is specified, and σ^2, variance of the subject responses. A proof of this property is included in Section 4.8.1. It is clear from our results that the benefits of group sequential analysis fall as r increases. The savings on the fixed sample test made when $r = 0$ are reduced by half around $r = 0.2$ or $r = 0.3$. However, delayed response tests do still offer substantial benefits for smaller r. For example, for $r = 0.1$, we retain approximately two-thirds of the benefits associated with a group sequential approach when response is immediate.

To illustrate how one should read Tables 4.2 - 4.5, we consider a comparative trial cited by Stallard & Todd (2003) used to test the efficacy of a new treatment for Alzheimer's disease. The primary endpoint is a subject's score on the Alzheimer's Disease Assessment Scale (ADAS) cognitive portion following twelve weeks of treatment, a response which is assumed to be normally distributed. We deviate slightly from the example of Stallard & Todd and suppose that the new treatment is to be compared only against control. Limited resources mean we can expect to recruit a maximum of $1.15n_{fix} = 120$ subjects. Accrual is expected to proceed at a constant rate of two subjects a week, so that in the absence of early stopping recruitment will be completed in $t_{max} = 60$ weeks. We assume that there will be no delay in processing responses ready for an interim analysis. In our notation, there is a delay of $\Delta_t = 12$ weeks in response and we are working under $r = \Delta_t/t_{max} = 0.2$. Let θ denote the effect size for

	r								
K	0	0.01	0.1	0.15	0.2	0.25	0.3	0.4	0.5
2	86.0	86.6	91.4	93.3	94.8	96.0	97.0	98.4	99.3
3	81.6	82.3	88.3	90.7	92.7	94.2	95.5	97.4	98.6
5	78.0	78.9	85.7	88.6	90.9	92.8	94.3	96.6	98.1
10	75.2	76.1	83.7	87.0	89.6	91.7	93.5	96.1	97.8
20	73.7	74.7	82.7	86.2	89.0	91.2	93.1	95.8	97.6

Table 4.2: Minima of F_1 expressed as a percentage of the corresponding fixed sample size for $\alpha = 0.05$, $\beta = 0.1$ and $R = 1.15$.

	r								
K	0	0.01	0.1	0.15	0.2	0.25	0.3	0.4	0.5
2	74.6	75.3	81.3	84.1	86.5	88.6	90.5	93.8	96.3
3	67.6	68.4	75.9	79.4	82.5	85.2	87.6	91.6	94.7
5	63.1	64.1	72.4	76.3	79.8	82.9	85.6	90.2	93.8
10	59.8	60.8	69.7	74.1	77.9	81.2	84.2	89.2	93.1
20	58.2	59.3	68.5	72.9	76.9	80.4	83.5	88.7	92.8

Table 4.3: Minima of F_2 expressed as a percentage of the corresponding fixed sample size for $\alpha = 0.05$, $\beta = 0.1$ and $R = 1.15$.

	r								
K	0	0.01	0.1	0.15	0.2	0.25	0.3	0.4	0.5
2	61.1	61.8	67.9	71.1	74.3	77.3	80.2	85.7	90.7
3	48.2	49.1	57.3	61.7	65.8	69.8	73.5	80.4	86.6
5	41.3	42.4	51.8	56.7	61.4	65.8	70.0	77.6	84.4
10	37.6	38.7	48.6	53.8	58.8	63.4	67.9	76.0	83.2
20	35.9	37.0	47.1	52.4	57.5	62.3	66.9	75.2	82.7

Table 4.4: Minima of F_3 expressed as a percentage of the corresponding fixed sample size for $\alpha = 0.05$, $\beta = 0.1$ and $R = 1.15$.

	r								
K	0	0.01	0.1	0.15	0.2	0.25	0.3	0.4	0.5
2	77.6	78.2	83.8	86.2	88.4	90.3	91.9	94.7	96.9
3	71.0	71.8	78.7	81.9	84.7	87.1	89.2	92.7	95.4
5	66.7	67.7	75.5	79.1	82.3	85.0	87.4	91.4	94.6
10	63.6	64.6	73.1	77.0	80.5	83.5	86.2	90.5	94.0
20	62.0	63.1	71.9	76.0	79.6	82.8	85.5	90.1	93.7

Table 4.5: Minima of F_4 expressed as a percentage of the corresponding fixed sample size for $\alpha = 0.05$, $\beta = 0.1$ and $R = 1.15$.

the new treatment. We want to test $H_0 : \theta \leq 0$ against $H_1 : \theta > 0$ using a $K = 5$-stage test with type I error probability 0.05 at $\theta = 0$ and power 0.9 at $\theta = 1$. If our test boundaries are chosen to minimise F_4, reading from Table 4.5 we see that the minimum of this objective function is 82.3, representing a saving of just under 18% on the fixed sample size. If instead the endpoint were defined to be ADAS score after six weeks, with all other parameters unchanged, we would have $r = 0.1$ and would make a saving of almost 25% on n_{fix}.

While the results given in Tables 4.2 - 4.5 are invariant to changes in either δ, σ^2 or the accrual rate c, varying these parameters will alter the set-up of the problem to be solved. For example, halving σ^2 will halve n_{max}, and so with the same recruitment rate, r is doubled. A similar note applies to varying δ. For example, consider the Alzheimer's trial example described above. If we want power 0.9 at $\theta = 2$, the fixed sample test requires a quarter of the number of subjects needed under $\delta = 1$. Hence, n_{max} is reduced from 120 to 30. If we still recruit two subjects a week, recruitment will now be completed after 15 weeks and so r increases to 0.8. In light of our results, we know that under this setting we can expect to make few savings on the sample size. Alternatively, slowing recruitment down to one subject a week would mean we would complete recruitment after 60 weeks as originally planned, and so r would remain unchanged at $r = 0.2$.

For any given r, as K increases, the efficiency gains associated with group sequential monitoring increase. However, the additional gains to be made by increasing the number of stages beyond $K = 5$ are minimal. For example, for $K = 20$ and $r = 0.2$, the minimum of F_4 is 79.6, representing a saving of just over 20% on the fixed sample test. However, when $K = 5$, we can still achieve approximately 80% of these gains so that the minimum of F_4 is 82.3. In practice, the logistical efforts required by interim monitoring are likely to prohibit more than five analyses being conducted (Choi & Lee, 1999).

Figure 4-1 plots how the efficiency gains on the fixed sample test offered by optimal two-stage tests vary as a function of r. For small values of r, there are still worthwhile gains to be made by adopting a group sequential approach. For example, when $r = 0.1$, tests minimising F_4 offer savings of over 15% on the fixed sample test. Beyond $r = 0.3$, there are few benefits remaining of interim monitoring for early stopping. We have found the same trends apply as α, β, K and R vary. Looking at tests minimising F_3, whose efficiency gains are most robust to increases in r, for $K = 2$ initial savings of 40% on n_{fix} made for immediate responses shrink by almost a half when $r = 0.3$. This is because a first interim analysis has to be scheduled between the time when data begin to accumulate and when recruitment closes. Even scheduling an analysis at the earliest

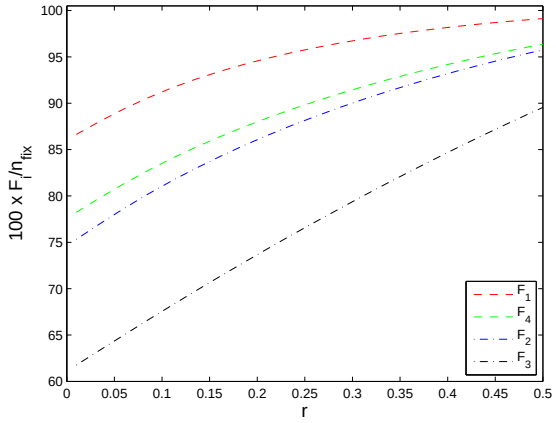


Figure 4-1: Objective functions achieved by the optimal delayed response GSTs, expressed as as a percentage of the fixed sample size, plotted against the delay parameter r. We fix $K = 2$, $R = 1.15$, $\alpha = 0.05$ and $\beta = 0.10$. Curves are labelled in the order that they appear in the figure.

opportunity, one is already committed to a large number of subjects in the pipeline.

Our results have important implications for how tests should be designed in practice. It is clear that we should try to minimise r as far as possible. Of course, a pertinent choice of primary endpoint is dictated by clinical considerations. However, there are other design choices within our control. For example, once the kth group of subjects have been observed, there is likely to be some delay while these data are cleaned before an analysis can take place, during which recruitment continues. The savings that can be made by speeding up this process may well be substantial and we look at this issue in more depth in the next section. Recruitment strategies can also be planned in light of the size of the delay in the subject response; recruiting subjects as quickly as possible may not be best for efficiency. Slowing down recruitment has obvious implications for the total length of the trial since one must balance the competing aims of achieving a low sample size and rapid time to a conclusion. One solution, proposed in Chapter 2 and explored further in Chapter 7, is to make measurements on a short-term endpoint and incorporate these into our test of $H_0 : \theta \leq 0$.

Tables 4.2 to 4.5 list results for values of the delay parameter for which tests can be found for all K listed, i.e., $r \leq 0.5$. When $r = 0.5$, almost none of the benefits for early stopping made when $r = 0$ remain. For example, the minima of $F_4 = 94.6\%$ of n_{fix} for a five-stage test. Hence, we lose little by restricting our attention to the case $r \leq 0.5$ since for higher values of r, it is unlikely that one would consider adopting a group sequential approach to achieve savings in sample size. Some interesting issues arise when looking at the behaviour of our optimal tests as r becomes large. For example,

is not possible to find tests satisfying our error constraints for sufficiently large values of r. In Section 4.8.2, we explain why this should occur and also discuss why it should be possible to find tests satisfying our error rate constraints with $\mathbb{E}(N;\theta) > n_{fix}$.

So far, we have measured the efficiency of the optimal delayed response tests derived in Chapter 3 using several criteria. In practice however, our choice of test is based on more than considerations of efficiency. In view of this, in the next section we use an illustrative example to explore our optimal tests further, examining what the boundaries of these tests look like and how they vary with r.

4.3 An example

Recall the trial into Alzheimer's described briefly in Section 4.2. Upon recruitment into the trial, subjects are allocated to either a new treatment or control. The superiority of the experimental treatment is to be tested based on subject scores on the ADAS cognitive portion following twelve weeks of treatment, a response which is assumed to be normally distributed. Let $X_{A,i} \sim N(\mu_A, \sigma^2)$ and $X_{B,i} \sim N(\mu_B, \sigma^2)$, $i = 1, 2, \ldots$, represent responses of subjects allocated to the experimental and control arms respectively. Define θ to be the effect size for the new treatment. We intend to use a $K = 3$ stage delayed response GST to test $H_0 : \theta \leq 0$ against $\theta > 0$ with type I error rate $\alpha = 0.05$ at $\theta = 0$ and power $1 - \beta = 0.9$ at $\theta = 1$. Limited resources mean that we can aim to recruit a maximum of $n_{max} = 1.15 n_{fix}$ subjects into the trial, where σ^2 is such that $n_{max} = 120$ subjects.

Recruitment will proceed at a constant rate of 2 subjects a week, with randomisation balanced in blocks of 2. Recruitment will be completed in $t_{max} = 60$ weeks. In our notation, $r = 12/60 = 0.2$, so that at each interim analysis, we will have a total of 24 subjects in the pipeline, 12 on each arm. Scheduling our interim analyses at times

$$t_k = r t_{max} + \frac{k}{K}(1 - r) t_{max} = 12 + \frac{k}{K} 48, \qquad \text{for } k = 1, \ldots, K - 1,$$

our three-stage test of $H_0 : \theta \leq 0$ generates the following sequence of the number of observed responses: $\{n_1 = 32, \tilde{n}_1 = 56, n_2 = 64, \tilde{n}_2 = 88, \tilde{n}_3 = 120\}$. However, our test will only proceed in this way if an interim analysis can be conducted the instant all necessary responses have been observed. Often in practice, this is not possible. Cleaning the interim analysis data set and, if necessary, arranging for a Data and Safety Monitoring Board to convene will all take time. Sooriyarachchi et al. (2003, p. 704) cite a stroke trial where it is anticipated that data transfer will take one month. However, throughout this period, recruitment will continue and new data accumulate.

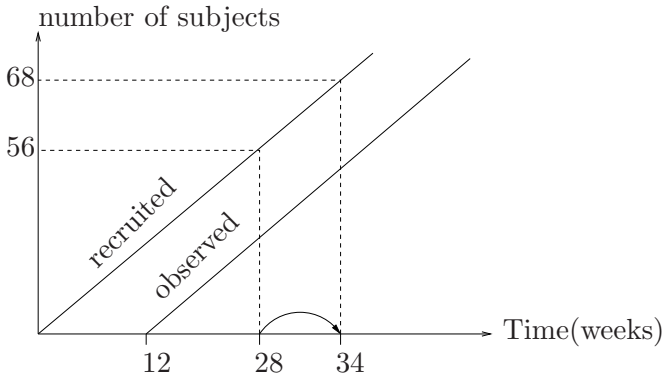


Figure 4-2: A schematic diagram of what will happen at the first interim analysis of the Alzheimer's trial when there is a delay of 6 weeks in data transfer for the first interim analysis.

Figure 4-2 illustrates how the Alzheimer's trial, planned assuming data transfer is immediate, proceeds if this in fact takes 6 weeks. The data set for the first interim analysis contains the observed responses collected in the first 28 weeks of the trial. In the 6 weeks it takes to clean and transfer these data, subjects continue to be admitted to the trial at a rate of 2 a week. Hence, by the time of the interim analysis, 68 subjects have been accrued and we have 36 subjects in the pipeline. Define $r' = (12 + 6)/t_{max} = 0.3$. Then, we could plan ahead for the delay in data transfer and, in preparation for interim analysis k, lock down the analysis data set at time

$$t_k^{'} = r' t_{max} + \frac{k}{K}(1 - r')t_{max} - 6 \qquad \text{for } k = 1, 2,$$

when $n_k = k(1 - r')n_{max}/3$ responses have been observed. Interim analysis k is then conducted at time $t_k = t_k^{'} + 6$ when $r' n_{max}$ subjects are in the pipeline; if termination is triggered, we wait for all $n_k + r' n_{max}$ responses to become available before conducting a decision analysis. Hence, the following sequence of sample sizes is generated: $\{n_1 = 28$, $\tilde{n}_1 = 64$, $n_2 = 56$, $\tilde{n}_2 = 92$, $\tilde{n}_3 = 120\}$. This is the same sequence as would be observed if the delay in the primary endpoint was 18 weeks, data transfer was immediate and analyses were scheduled following the pattern (2.3); the same test will be optimal for F_i in both cases. Hence, we can read off results for the scenario that $\Delta_t = 12$ weeks and data transfer takes 6 weeks by looking at entries for $r = 0.3$ in Tables 4.2 - 4.5 which were derived assuming immediate data transfer. Similarly, if data transfer takes 12 weeks, planning ahead for this delay, the GST generates the sequence of sample sizes $\{n_1 = 24$, $\tilde{n}_1 = 72$, $n_2 = 48$, $\tilde{n}_2 = 96$, $\tilde{n}_3 = 120\}$. We can read off results for this case by looking at results for $r = 0.4$ when there is no delay for data transfer.

We know from Section 4.2, that in terms of minimising expected sample size, we do

r	l_1	u_1	c_1	l_2	u_2	c_2	c_3
0	−0.013	2.215		0.779	2.088		1.733
0.2	−0.249	2.241	1.307	0.495	2.072	1.532	1.730
0.3	−0.308	2.183	1.367	0.412	1.996	1.559	1.731
0.4	−0.322	2.078	1.426	0.357	1.884	1.586	1.732

Table 4.6: Boundaries of three-stage delayed response GSTs of $H_0 : \theta \leq 0$ against $\theta > 0$ minimising F_2. Tests optimised under $\alpha = 0.05$, $\beta = 0.1$, $R = 1.15$ and $\delta = 1$. Interim analysis $k = 1, 2$ is scheduled at information level $I_k = k(1 - r)I_{max}/3$ and tests terminate at $\tilde{I}_3 = I_{max}$. Entries for $r = 0$ correspond to the optimal standard design when there is no delay in response.

better to keep to a minimum the number of subjects in the pipeline at an interim analysis. For our Alzheimer's trial, the savings in sample size which can be made by speeding up the cleaning process may well be worth the additional investment it necessitates. Referring to the entries in Table 4.3 for $r = 0.4$, $r = 0.3$ and $r = 0.25$, we see that reducing the time for data transfer from 12 weeks to 6 weeks means we can save, for F_2, 4% more on n_{fix}. Reducing this delay even further to 3 weeks, we save another 2.4%. Table 4.6 lists the critical values depicted in Figure 4-3 of an optimal three-stage test minimising F_2 when we anticipate no delay for data transfer ($r = 0.2$), a delay of 6 weeks ($r = 0.3$) and a delay of 12 weeks ($r = 0.4$) in our Alzheimer's trial. For reference, we also plot the boundaries of the optimal standard GST minimising F_2 with analyses scheduled at equally spaced information levels between 0 and I_{max}. For $r = 0.2$, the upper boundaries do not change greatly from those of the standard GST. However, the continuation region at each interim analysis does widen, reflecting the fact that as r increases, I_k and $I_k/\tilde{I}_k$ both decrease, for $k = 1, \ldots, K - 1$.

From now on, we make the simplifying assumption that data transfer is immediate, knowing that if it is not, one can read off the results for this case by looking at the appropriate value of the delay parameter r.

4.4 Comparison with a group sequential test designed for immediate responses

So far in this thesis, we have followed the same coherent design approach to dealing with delayed responses, following designs of the form (2.2). In contrast, an alternative approach is to design a standard GST using the data available at each analysis. These designs ignore the data in the pipeline at interim points; not using it after termination either. Practising statisticians may want to see evidence of large potential savings before changing practice from using standard tests, which they have experience of using

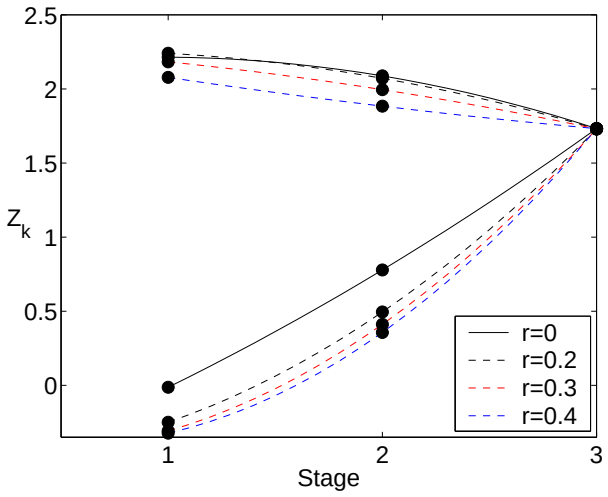


Figure 4-3: Boundaries at interim analyses $k = 1, 2$ and decision analysis $K = 3$ of optimal three-stage tests of $H_0 : \theta \leq 0$ against $\theta > 0$. Tests minimise F_2 and are found under $\alpha = 0.05$, $\beta = 0.1$ and $R = 1.15$. Boundaries for the case $r = 0$ correspond to the optimal standard GST. We have interpolated through these boundaries to highlight the shape of the continuation regions.

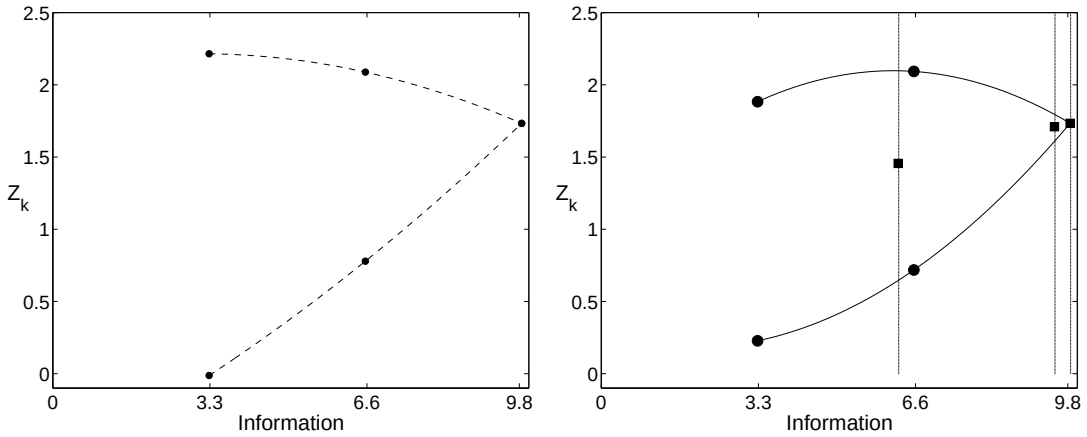
and are advocated by regulators (see "E9 Statistical Principles for Clinical Trials", (1998)). In this section, we compare the two approaches and calculate the savings that can be made by planning ahead for the delay in our response.

Before making any comparisons, we first explain in more detail how one might apply a K-stage standard GST of $H_0 : \theta \leq 0$ against $\theta > 0$ with delayed data. Suppose the standard design is formulated assuming equal group sizes, so that interim analysis $k = 1, \ldots, K-1$ is scheduled once $n_k = kn_{max}/K$ subjects have been observed, or equivalently once information $I_k = kI_{max}/K$ has been accrued. For consistency with previous notation, we say the final analysis is conducted at information level $\tilde{I}_K = I_{max}$. This scheduling is the same for all values of the delay parameter r. The standard test is applied using only those data available at the interim analysis: the test does not wait for the pipeline observations before rejecting or accepting H_0. The boundaries of the standard test can be chosen to minimise expected sample size, only counting those subjects observed at the time of termination. Let O_T represent the number of subjects observed at time of termination and define

$$F_2^\star = 0.5\{\mathbb{E}(O_T; \theta = 0) + \mathbb{E}(O_T; \theta = \delta)\}.$$

Figure 4-4(a) depicts the standard GST minimising $F_2^\star$ when $K = 3$.

Delayed response GSTs are planned to take account of the subjects in the pipeline at



(a) Optimal standard GST

(b) Optimal delayed response GST

Figure 4-4: Boundaries defining a delayed response and standard GST for $K = 3$, $r = 0.3$, $\alpha = 0.05$, $\beta = 0.1$, $R = 1.15$ and $\delta = 1$. The delayed response test minimises F_2 for the information sequence $\{I_1 = 3.3, \tilde{I}_1 = 6.2, I_2 = 6.6, \tilde{I}_2 = 9.5, \tilde{I}_3 = 9.8\}$ and the standard test minimises $F_2^\star$ for the sequence $\{I_1 = 3.3, \tilde{I}_1 = 6.2, I_2 = 6.6, \tilde{I}_2 = 9.5, \tilde{I}_3 = 9.8\}$.

each interim analysis. If recruitment is closed, we wait to observe all rn_{max} pipeline subjects before deciding whether to reject H_0 and expected sample sizes include all admitted subjects. Figure 4-4(b) shows the boundaries of our optimal three-stage delayed response test minimising F_2 when $r = 0.3$ and interim analyses are timed at information levels

$$I_k = \frac{k}{K} I_{max}, \qquad \text{for } k = 1, \ldots, K-1.$$

Comparing the shape of the continuation regions of the two tests at each interim analysis, we see how the delayed response test takes calculated decisions; the fact that more data will be available later to help correct a "mistake" could be behind the narrower boundary at I_1 for the delayed response GST.

Even though the standard test stipulates that we should not wait for the rn_{max} pipeline subjects to respond before making a decision at an interim analysis, evaluating these standard designs properly means the expected sample size ought to include all admitted subjects. Table 4.7 lists values of F_2 attained by optimal delayed response and standard GSTs for several values of r. Scheduling our interim analyses so that $I_k = kI_{max}/K$, for $k = 1, \ldots, K-1$, means that for $r > 0.2$, recruitment will be completed before some interim analyses can take place. If this happens, we forfeit any future interim analyses and wait for all n_{max} subjects to be observed before conducting a final analysis. The boundaries of the delayed response GSTs adapt to this constraint so that the tests based on $K^\star < K$ stages continue to satisfy their error constraints. However, the boundaries of the standard GSTs remain fixed.

r	Standard GST	Delayed response GST
0.01	64.1	64.1
0.1	73.0	72.7
0.2	83.0	80.8
0.3	91.3	86.7
0.4	99.5	91.6

Table 4.7: Attained values of F_2 expressed as a percentage of the corresponding fixed sample size for standard tests minimising $F_2^\star$ and delayed response tests minimising F_2. For both tests, interim analyses are timed at information levels $I_k = kI_{max}/K$, for $k = 1, \ldots, K-1$. Tests are found under $\alpha = 0.05$, $\beta = 0.1$, $R = 1.15$ and $K = 5$.

For small r, the standard and delayed response tests differ little in their performances. This is to be expected since at any interim analysis, the amount of information in the pipeline will be a small proportion of the total observed. Hence, it contributes little to our decision making process. For larger r, the performances of the standard and delayed response tests diverge. For $r = 0.3$, our savings on the fixed sample test for F_2 increase from 8.7% for a standard design to 13.3% for a delayed response test, i.e., we can save almost 5% more on the fixed sample test by providing a proper treatment of the pipeline data. For a Phase III study, this will amount to a substantial saving. However, any additional gains made by using our delayed response GSTs are only worthwhile if a group sequential approach is viable. Indeed, as r increases beyond 0.4, the advantages of allowing our test boundaries to vary with r is clear, although the benefits for early stopping associated with either group sequential approach are small.

So far, we have not considered what effect the timing of our interim analyses has on the efficiency of our tests. When adopting a group sequential approach with delayed responses, it is not obvious how we should schedule them. When r is large, recruitment is well under way by the time data begin to accumulate. It is hard to find a time with enough useful information on which to base a decision at the interim analysis and still have scope for saving subject numbers. For example, consider a two-stage delayed response test designed under $R = 1.15$ and $r = 0.2$. Indexing the timing of the interim analysis by the ratio I_1/I_{fix}, Table 4.8 lists the performances of optimal delayed response GSTs for F_2 as I_1/I_{fix} varies. When optimising with respect to I_1/I_{fix}, a constrained minimisation algorithm is used to ensure this ratio lies in the interval $(0, (1-r)R)$, i.e., so that the interim analysis occurs before recruitment is completed.

Looking at the entry for $I_1/I_{fix} = 0.575$, we see that if we time the interim analysis for a delayed response GST with $K = 2$ and $r = 0.2$ to be based on information

I_1/I_{fix}	Minima of $100(F_2/n_{fix})$
0.1	100.7
0.2	94.0
0.3	89.1
$0.438^\star$	86.4
0.46	86.5
0.575	88.9
0.7	95.2

Table 4.8: Values of F_2 expressed as a percentage of n_{fix} achieved by optimal two-stage GSTs as the timing of the first interim analysis varies. The superscript $\star$ indicates the optimal timing. Tests are optimised under $K = 2$, $\alpha = 0.05$, $\beta = 0.1$, $R = 1.15$ and $r = 0.2$.

$I_1 = I_{max}/2$ we make a saving of 11% on the fixed sample test for F_2. The results presented in Section 4.2 are for optimal tests derived under the timing schedule (2.3): interim analysis k is based on information $I_k = k(1-r)I_{max}/K$, for $k = 1, \ldots, K-1$. Looking at the entry for $I_1/I_{fix} = 0.46$, we see that for a two-stage test, this schedule is close to optimal. It is clear that the efficiency of our tests can be increased by adapting the timings of analyses to the delay in the system. Table 4.9 compares values of F_2 attained by optimal $K = 5$-stage delayed response GSTs under information schedule (2.4) with those attained by standard GSTs minimising $F_2^\star$ with equally spaced interim analyses. For $r = 0.4$, following a standard group sequential approach means we make a saving of less than 1% on the fixed sample test with a five-stage test. This increases to just under 10% when conducting a delayed response GST, allowing the timings of the interim analyses to vary with r. We see that changing our design approach has meant a group sequential approach is now viable.

4.5 Behaviour of optimal test boundaries

Following our new test structure means that we will be uncertain of whether we will eventually reject or accept H_0 at the crucial point when deciding to stop recruitment. One might expect the direction in which the test statistic sample path exits the continuation region to be a good indicator of the hypothesis decision which will be made at the decision analysis. However, there may be a "reversal" with the eventual decision being contrary to that anticipated at the interim analysis. We say a positive to negative reversal occurs at stage k if we are prompted to terminate recruitment at the interim analysis by observing $Z_k \geq u_k$ but fail to reject H_0 at the decision analysis. Conversely, should we reject H_0 after terminating recruitment with $Z_k \leq l_k$, we say a negative to positive reversal has occurred. Terminating recruitment is an important step in any clinical trial. Tests of the form (2.2) stipulate that by terminating recruitment, we

r	Standard GST	Delayed response GST
0.01	64.1	64.0
0.1	73.0	72.4
0.2	83.0	79.8
0.3	91.3	85.6
0.4	99.5	90.2

Table 4.9: Attained values of F_2 expressed as a percentage of the corresponding fixed sample size for standard tests minimising $F_2^\star$ and delayed response tests minimising F_2. For standard designs, interim analyses are timed at information levels $I_k = kI_{max}/K$, for $k = 1, \ldots, K-1$, while for delayed response tests, they are scheduled at levels $I_k = k(1-r)I_{max}/K$. Tests are found under $\alpha = 0.05$, $\beta = 0.1$, $R = 1.15$ and $K = 5$.

commit to observing only those subjects currently recruited, even if at the decision analysis we would ideally like to continue sampling. Sooriyarachchi et al. (2003) note that a particular regret would be terminating recruitment early based on a positive trend only to then narrowly miss significance at the decision analysis; sponsors may be loath to close recruitment early if there is a large risk of this occurring. Clearly, a test's reversal probabilities are of interest to us and in this section we calculate these probabilities for the optimal delayed response tests derived in Section 4.2.

Let us consider a K-stage delayed response GST. For each $k = 1, \ldots, K-1$, define

$$\psi_k(\mu) = \mathbb{P}(l_1 < Z_1 < u_1, \ldots, l_{k-1} < Z_{k-1} < u_{k-1}, Z_k \geq u_k, \tilde{Z}_k < c_k; \theta = \mu)$$
$$\xi_k(\mu) = \mathbb{P}(l_1 < Z_1 < u_1, \ldots, l_{k-1} < Z_{k-1} < u_{k-1}, Z_k \leq l_k, \tilde{Z}_k \geq c_k; \theta = \mu),$$

where $\psi_k(\mu)$ and $\xi_k(\mu)$ are the stage k reversal probabilities under $\theta = \mu$. Note that $\psi_k(0)$ and $\xi_k(\delta)$ are probabilities that a reversal at stage k leads us to make the correct decision at the decision analysis, i.e., the pipeline data help us switch to the correct decision. Conversely, $\psi_k(\delta)$ and $\xi_k(0)$ are probabilities that a reversal at stage k leads us to incorrectly accept or reject H_0 respectively. Let $\lambda_k(z_k; \theta)$ be the conditional probability that we reject H_0 at stage k given $Z_k = z_k$. Tables 4.10 and 4.11 list reversal and conditional rejection probabilities for a two-stage test minimising F_4.

We see that for small r, the probability that a reversal occurs at the first stage is almost negligible. Most notably, if the true effect size for our new treatment is δ, the probability that we accept H_0 after stopping at the first interim analysis anticipating rejection is close to 0. Looking at the conditional rejection probabilities, we also see that when r is small, in effect our hypothesis decision is made on the basis of those data available at the interim analysis. Given we exit via the upper boundary of the continuation region we will almost surely reject H_0 at the decision analysis and vice

	Switch to correct decision		Switch to incorrect decision	
r	$\psi_1(0)$	$\xi_1(\delta)$	$\xi_1(0)$	$\psi_1(\delta)$
0.01	8×10^{-9}	1×10^{-9}	5×10^{-10}	1×10^{-8}
0.1	0.0015	0.0019	0.0008	0.0016
0.2	0.0057	0.0087	0.0028	0.0048
0.3	0.0121	0.0194	0.0050	0.0080
0.4	0.0222	0.0362	0.0074	0.0114
0.5	0.0395	0.0642	0.0104	0.0156

Table 4.10: Reversal probabilities for two-stage delayed response GSTs minimising F_4 under $\alpha = 0.05$, $\beta = 0.10$ and $R = 1.15$.

versa. The pipeline data are not helping us make better decisions and the rn_{max} subjects are acting as a "fixed penalty" in terms of efficiency which increases linearly with r. Hence, we infer that for small r, as r increases, the benefits for early stopping associated with GSTs will decrease linearly. This trend is apparent in Figure 4-1, for all objective functions considered.

For larger values of r, the pipeline data play a greater role in the decision making process. The probability that they cause us to switch to the right decision at the decision analysis increases. For example, the conditional probability under $\theta = \delta$ that we reject H_0 at stage k given $Z_k = l_k$ increases from 0.301 to 0.713 as r increases from 0.2 to 0.5. Our reversal probabilities also increase. Under $r = 0.5$, the probability of a positive to negative reversal at the first stage under $\theta = 0$ reaches 0.04 while under $\theta = \delta$, the probability of a negative to positive reversal is 0.06. We infer that our decision of whether to reject or accept H_0 is no longer made using only those data available at the interim analysis. Instead, at the interim analysis we can think of these data as helping us choose the next group size. For example, consider a two-stage delayed response test. At the first analysis, using the n_1 observed responses we must choose between continuing sampling and terminating recruitment, i.e. do we collect $\tilde{n}_1 - n_1$ or $\tilde{n}_2 - n_1$ additional responses? There are clear parallels between this nonadaptive GST and a two-stage adaptive test, where the second stage group size is adapted to first stage data in a pre-planned way. Note that the adaptive test has the added flexibility however, of having a fully variable second group size which can be allowed to vary over a continuum of values. We compare the nonadaptive and adaptive designs in more detail to see the gains that can be made by testing adaptively.

r	$\lambda_1(u_1;\theta=0)$	$\lambda_1(u_1;\theta=\delta)$	$\lambda_1(l_1;\theta=0)$	$\lambda_1(l_1;\theta=\delta)$
0.01	1.000	1.000	$6.41\,\mathrm{x}\,10^{-8}$	$3.39\,\mathrm{x}\,10^{-7}$
0.1	0.842	0.977	0.0132	0.110
0.2	0.620	0.956	0.0272	0.301
0.3	0.448	0.944	0.0357	0.467
0.4	0.315	0.934	0.0424	0.603
0.5	0.216	0.924	0.0487	0.713

Table 4.11: Conditional rejection probabilities for two-stage delayed response GSTs minimising F_4 under $\alpha = 0.05$, $\beta = 0.10$ and $R = 1.15$.

4.6 Adaptive sampling rules

Table 4.12 lists the minima of F_4 expressed as a percentage of n_{fix} attained by optimal two-stage adaptive designs and nonadaptive delayed response GSTs. The adaptive designs are examples of the "sequentially planned sequential tests" proposed by Schmitz (1993); they are optimal in the class of tests where group sizes are adapted to observed responses in a pre-planned way. Trial sponsors will know ahead of time how they are required to adapt to the observed data in every possible scenario. The interim analysis for the adaptive and nonadaptive tests is timed at the same information level. For comparison with the nonadaptive tests, we constrain the second stage group size for the adaptive test to lie in the interval $[\tilde{n}_1 - n_1, n_{max} - n_1]$, where all tests are found under $r = 0.2$.

The class of adaptive designs must contain nonadaptive designs as special cases. After all, using an adaptive design one can always choose to not adapt at an interim analysis. Hence, optimal adaptive designs do perform better than their nonadaptive counterparts. However, our conclusion from Table 4.12 is that there is not much of an advantage to having the fully variable second group size. For most values of I_1/I_{fix}, the average $\mathbb{E}(N;\theta)$ for the best nonadaptive test is within 1% of n_{fix} of the optimal adaptive test's average expected sample size. As the information ratio gets small, the benefits of an adaptive approach do increase although they still remain small. These conclusions are in agreement with the findings of Jennison & Turnbull (2006), who compare optimal "sequentially planned sequential tests" minimising F_4 with optimal standard GSTs when response is immediate. They find that for most values of K and inflation factor R, the average expected sample size for the optimal standard GST is within 2% of n_{fix} of the optimal adaptive design.

Faldum & Hommel (2007) propose an adaptive design approach for incorporating pipeline observations based on the two-stage procedures of Bauer & Köhne (1994). The first stage is planned with power $1 - \beta_1$ at the alternative $\theta = \delta_1$. Let P_1 be the

I_1/I_{fix}	Optimal delayed response	Optimal adaptive
0.1	100.7	97.9
0.2	94.5	93.3
0.3	90.3	89.6
$0.425^{\dagger}$	88.3	87.9
$0.431^{\ddagger}$	88.3	87.9
0.6	91.4	91.2
0.7	95.9	95.7

Table 4.12: Minima of F_4 expressed as a percentage of n_{fix} achieved by optimal two-stage delayed response and adaptive GSTs as the timing of the interim analysis varies. Tests are found under $\alpha = 0.05$, $\beta = 0.1$, $R = 1.15$ and $r = 0.2$. Results for optimal adaptive designs were computed by Jennison (2006). Information ratios labelled with superscripts † and ‡ are optimal for the adaptive and nonadaptive designs, respectively.

p-value based on the first n_1 responses and let P_a denote the p-value based on the responses of the $\tilde{n}_1 - n_1$ subjects in the pipeline at the interim analysis. At the design stage, we must specify a conditional error function $\mathcal{C}(p_1)$ defined as

$$\mathcal{C}(p_1) = P_{\theta=0}(\text{Reject } H_0 | P_1 = p_1).$$

The conditional error function is defined so that the test has overall type I error rate α under $\theta = 0$, i.e.,

$$\int_0^1 \mathcal{C}(p_1) dp_1 = \alpha.$$

Recruitment stops at the first interim analysis if $p_1 \leq \alpha_0$ or $p_1 \geq \alpha_1$, otherwise sampling continues. If stopping occurs with $p_1 \geq \alpha_1$, H_0 is accepted without waiting for the pipeline observations. However, in the case of stopping recruitment, H_0 can only be rejected if $p_1 \leq \alpha_0$ and this positive result is repeated in the overrun data with $p_a \leq \mathcal{C}(p_1)$. If sampling is to continue, the second stage sample size is planned to attain conditional power $1 - \beta_2$ at $\delta(p_1)$. At the final analysis, H_0 is rejected if $p_2 \leq \mathcal{C}(p_1)$.

Jennison & Turnbull (2006, p. 13) note that adaptive tests which base inference on a non-sufficient statistic and modify sample size according to a conditional power criterion are likely to be inefficient. The authors give examples when the response is immediate, where such adaptive tests can be beaten everywhere in terms of efficiency by a nonadaptive group sequential design with a matched power curve. Based on this reasoning and our results in Table 4.12, we conclude that it is important to formulate nonadaptive schemes which can deal efficiently and systematically with delayed responses; few gains on our optimal delayed response designs can be made by testing adaptively, even if we act optimally.

4.7 Discussion

We have seen that testing group sequentially continues to deliver savings in expected sample size when there is a delay in response. However, these benefits are reduced from the values seen when response is immediate, becoming unconvincing as r increases to 0.3 and beyond. One question of practical interest is in what circumstances might it be appropriate to use an ordinary GST, despite ignoring the pipeline data. Our response looks at two key issues: efficiency and interpretability. Addressing the first point, we find that certainly for small $r \leq 0.01$ there is little to gain in terms of sample size by switching from an ordinary GST to one of our new designs. Indeed, Tables 4.7 and 4.9 show that $\mathbb{E}(N;\theta)$ increases by around rn_{max} in both cases. This is also approximately true for $r \leq 0.1$. However, for larger values of r there are clear benefits to adopting a group sequential design which provides a proper treatment of the pipeline data. Looking at Table 4.7, we see that for $K = 5$ and $r = 0.3$ we can save an additional 5% on n_{fix} for F_2 by using a GST designed for delayed responses, a considerable saving in the context of a Phase III study. Our new designs are subtle in their use of the pipeline data. Tables 4.10 and 4.11 show that for small r, the probability of "switching" decision at a decision analysis is very small. In effect it is as if our tests are deciding whether to reject or accept H_0 without waiting for the pipeline data to become available. For larger values of r however, this is certainly not the case. We find that delayed response GSTs take calculated decisions at an interim analysis, taking into account the pipeline data to come later and allowing some probability of switching at the decision analysis.

Even if there are no objections for reasons of efficiency to using an ordinary GST ignoring the pipeline data, issues of interpretability could still arise. For example, suppose one implements a standard GST, crossing the upper boundary at interim analysis k with $Z_k = 2.5$. However, once the pipeline data come in the Z-statistic falls to $\tilde{Z}_k = 2.2$; if there is an obligation to report the follow-up data, the standard GST breaks down as it is not clear whether a positive result can still be claimed. One strategy could be to reject H_0 only if $\tilde{Z}_k \geq u_k$, although it is clear the power of the test will be reduced since the decision constants c_k, $k = 1, \ldots, K$ for our optimal delayed response tests are usually someway lower than u_k. In contrast to the problems that arise for standard GSTs in this situation, our delayed response designs are fine. Scenarios such as this were envisaged at the design stage and we make decisions according to a pre-specified rule chosen to ensure the overall type I error rate and power of our test are as required.

We conclude that there are compelling reasons for switching from using standard GSTs to our new designs for delayed responses. With this in mind, in the next chapter we

derive methods of inference for on termination of delayed response GSTs.

4.8 Appendix

4.8.1 Proof of invariance property

We wish to compare two treatments A and B. Let $X_{A,i} \sim N(\mu_A, \sigma^2)$ and $X_{B,i} \sim N(\mu_B, \sigma^2)$, $i = 1, 2, \ldots$, represent responses of subjects allocated to A and B, respectively. For delay parameter r, we wish to design a K-stage delayed response GST of $H_0 : \theta \leq 0$ against $\theta > 0$ of size α and power $1 - \beta$ at $\theta = \delta$ with maximum information level $I_{max} = RI_{fix}$. Patient accrual is to be divided equally between treatments, so that the test requires a maximum total sample size of $n_{max} = 4\sigma^2 I_{max}$ subjects.

Theorem 2. *For this problem, fixing r, K, α, β and R, the values of objective functions F_i, $i = 1, \ldots, 4$, expressed as a percentage of the corresponding fixed sample size are invariant to changes in σ^2 and δ.*

Proof: We prove this claim by considering two different versions of the testing problem outlined above. For the first problem, suppose responses are distributed with variance σ_1^2 and we wish to specify power at the alternative $\theta = \delta_1$. Then, the test of H_0 for this problem has maximum information level

$$I_{1,max} = R\frac{\{\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)\}^2}{\delta_1^2}.$$

For $k = 1, \ldots, K-1$, let $I_{1,k}$ denote the information at interim analysis k and for each $k = 1, \ldots, K$, let $\tilde{I}_{1,k}$ denote the information at decision analysis k. Analyses are scheduled at information levels

$$I_{1,k} = \frac{k}{K}(1-r)I_{1,max} \qquad \tilde{I}_{1,k} = I_{1,k} + rI_{1,max}, \quad \text{for } k = 1, \ldots, K-1,$$

and $\tilde{I}_K = I_{1,max}$. Suppose we order these information levels from smallest to largest, where $I_{1,(k)}$ denotes the kth smallest and X_k the associated standardised test statistic. Then, given information levels $\{I_{1,(1)}, \ldots, I_{1,(2K-1)}\}$, the sequence $\{X_1, \ldots, X_{2K-1}\}$ has joint distribution

(i) $(X_1, \ldots, X_{2K-1})$ are jointly multivariate normal,

(ii) $X_k \sim N(\theta\sqrt{I_{1,(k)}}, 1)$, for $k = 1, \ldots, 2K-1$,

(iii) $Cov(X_{k_1}, X_{k_2}) = \sqrt{I_{1,(k_1)}/I_{1,(k_2)}}$, for $1 \leq k_1 \leq k_2 \leq 2K-1$.

Let

$$\mathcal{C}_1 = \{l_k, u_k, c_j, k = 1, \ldots, K-1, j = 1, \ldots, K\}$$

be the set of boundary constants for monitoring the $\{Z_k, \tilde{Z}_k\}$ which define an appropriate test of H_0 for this problem.

Consider a second variant on our original testing problem where responses are distributed with variance σ_2^2 and power is specified at $\theta = \delta_2$. The maximum information level of the GST for this second problem is $I_{2,max} = (\delta_1/\delta_2)^2 I_{1,max}$. Analyses for this second GST are scheduled following the same pattern as for problem 1. Ordering the information sequence thus generated from smallest to largest, we have $\{I_{2,(k)} = (\delta_1/\delta_2)^2 I_{1,(k)}\ ;\ k = 1, \ldots, 2K-1\ \}$. Given these information levels, the associated sequence of ordered standardised test statistics $\{Y_1, \ldots, Y_{2K-1}\}$ has joint distribution

(i) $(Y_1, \ldots, Y_{2K-1})$ are jointly multivariate normal,

(ii) $Y_k \sim N((\delta_1/\delta_2)\theta\sqrt{I_{1,(k)}}, 1)$, for $k = 1, \ldots, 2K-1$,

(iii) $Cov(Y_{k_1}, Y_{k_2}) = \sqrt{I_{1,(k_1)}/I_{1,(k_2)}}$, for $1 \leq k_1 \leq k_2 \leq 2K-1$.

The joint distribution under $\theta = \mu$ of $\{Y_1, \ldots, Y_{2K-1}\}$ given $\{I_{2,(k)}\colon\ k = 1, \ldots, 2K-1\}$ is the same as the distribution under $\theta = \mu(\delta_1/\delta_2)$ of $\{X_1, \ldots, X_{2K-1}\}$ given $\{I_{1,(k)}\colon\ k = 1, \ldots, 2K-1\}$. Then, corresponding tests of H_0 for problems 1 and 2 are based on information sequences and sets of boundary constants $\{I_{1,(k)}\}$ and $\mathcal{C}_1$ and $\{(\delta_1/\delta_2)^2 I_{1,(k)}\}$ and $\mathcal{C}_1$. We refer to these as tests 1 and 2, respectively, where the stopping probabilities for test 2 under $\theta = \mu$ are equal to the probabilities for test 1 under $\theta = \mu(\delta_1/\delta_2)$. Define N_1 and $I_{T,1}$ to be the number of subjects recruited and information for θ on termination of test 1. Define N_2 and $I_{T,2}$ similarly for test 2. We can write $\mathbb{E}(N_1; \theta) = 4\sigma_1^2\, \mathbb{E}(I_{T,1}; \theta)$ and $\mathbb{E}(N_2; \theta) = 4\sigma_2^2\, \mathbb{E}(I_{T,2}; \theta)$. It follows that

$$\frac{\mathbb{E}(N_2; \theta = \mu)}{\mathbb{E}(N_1; \theta = \mu(\delta_1/\delta_2))} = \left(\frac{\sigma_2\delta_1}{\sigma_1\delta_2}\right)^2. \tag{4.1}$$

Let $n_{fix}(\alpha, \beta, \sigma^2, \delta)$ denote the fixed sample size required to test $H_0 : \theta \leq 0$ against $\theta = 0$ with type I error rate α at $\theta = 0$, power $1-\beta$ at $\theta = \delta$ when responses are distributed with variance σ^2. It is clear that $n_{fix}(\alpha, \beta, \sigma_2^2, \delta_2)/n_{fix}(\alpha, \beta, \sigma_1^2, \delta_1)$ is equal to the right hand side of (4.1). It follows that

$$\frac{\mathbb{E}(N_2; \theta = \mu)}{n_{fix}(\alpha, \beta, \sigma_2^2, \delta_2)} = \frac{\mathbb{E}(N_1; \theta = \mu(\delta_1/\delta_2))}{n_{fix}(\alpha, \beta, \sigma_1^2, \delta_1)}, \tag{4.2}$$

and therefore values of F_i, $i = 1, \ldots, 3$ expressed as a percentage of the corresponding fixed sample size are invariant to changes in δ and σ^2. For objective function F_4,

applying equation (4.2) we obtain

$$\frac{\int_\Theta \mathbb{E}(N_2;\mu)\frac{2}{\delta_2}\phi\left(\frac{\mu-\delta_2/2}{\delta_2/2}\right)d\mu}{n_{fix}(\alpha,\beta,\sigma_2^2,\delta_2)} = \frac{\int_\Theta \mathbb{E}(N_1;\mu(\delta_1/\delta_2))\frac{2}{\delta_2}\phi\left(\frac{\mu-\delta_2/2}{\delta_2/2}\right)d\mu}{n_{fix}(\alpha,\beta,\sigma_1^2,\delta_1)}. \tag{4.3}$$

Applying the substitution $x = \mu(\delta_1/\delta_2)$ to the integral in the numerator of the right hand side of (4.3) we find that values of F_4 expressed as a percentage of the corresponding fixed sample size are invariant to changes in δ and σ^2. □

4.8.2 Properties of group sequential tests as r gets large

For large values of r, it is not possible to find tests with error rates $\alpha^\star = \alpha$ and $\beta^\star = \beta$. To explain this, note that as r becomes large, it is difficult to find suitable timings for our analyses. By the time of the first interim analysis we need some responses to be available, but then nearly all n_{max} subjects are recruited and $\tilde{I}_1 > I_{fix}$. Hence, even if we always stop at the first interim analysis the test's power will exceed $1-\beta$ at $\theta = \delta$. Tables 4.2 to 4.5 list results for values of the delay parameter for which tests can be found for all K listed. Where a test exists for $r > 0.5$, the minima of F_i, $i = 1,\ldots,4$, follow the same trends as r increases. For some values of r, there exist optimal tests satisfying our error constraints with $\mathbb{E}(N;\theta) > n_{fix}$. For example, operating under $K = 2$, $R = 1.15$ and $r = 0.7$, at the first interim analysis we are in effect choosing between making our final hypothesis decision based on information level $\tilde{I}_1 = 0.98\, I_{fix}$ and $\tilde{I}_2 = 1.15\, I_{fix}$. Our power requirement dictates that we cannot always choose to close recruitment at the earliest opportunity: the fixed sample test of level α based on information $\tilde{I}_1$ will fail to attain power $1-\beta$ at $\theta = \delta$. Hence, we must choose information level $\tilde{I}_2 > n_{fix}$ at least some of the time and it can be shown that, in this case, the minima of $F_1 = 100.02\, n_{fix}$.

At first glance, it is perhaps surprising that a test with $\mathbb{E}(N;\theta) > n_{fix}$ does not have power exceeding $1-\beta$ at $\theta = \delta$. In order to explain how this is possible, we look at the power attained by a two-stage delayed response GST of H_0 in the limiting case when $r = 1+\epsilon$, $\epsilon > 0$. In this case, recruitment will be completed before we observe any data. Hence, at the interim analysis, we must decide whether to base our hypothesis decision on either $\tilde{n}_1$ or $\tilde{n}_2$ observations according to a random decision rule, such as tossing a coin, which is independent of the data. In this limiting case, our test of H_0 can be thought of as a fixed sample test based on a random sample size N. Let $P(N;\theta=\delta)$ denote the conditional power of this fixed sample test under $\theta = \delta$ given N, where $P(n_{fix};\theta=\delta) = 1-\beta$. The overall power of the test at this alternative is given by $\mathbb{E}(P(N;\theta=\delta))$. Following the arguments of Jennison & Turnbull (2003, Section 3.5), we note that $P(N;\theta=\delta)$ will be concave in N for values for which $P(N;\theta=\delta) > 0.5$.

Let $\tilde{n}_1$ and $\tilde{n}_2$ satisfy this condition. Applying Jensen's inequality, we obtain

$$\mathbb{E}(P(N;\theta=\delta)) \leq P(\mathbb{E}(N);\theta=\delta),$$

and see that the overall power of the test is bounded above by the power attained given $N = \mathbb{E}(N)$. Hence, under $r = 1 + \epsilon$, we see that it is possible for a fixed sample test with random sample size to have overall power $1 - \beta$ at $\theta = \delta$ when $\mathbb{E}(N) > n_{fix}$.

A single stage test with randomly generated sample size is not efficient, although we are getting something close to this situation with our delayed response tests as r becomes large. Hence, based on the reasoning given above, we claim that it is reasonable that for large r we should have delayed response tests satisfying our error constraints with $\mathbb{E}(N;\theta) > n_{fix}$ for general K.

Chapter 5

Inference on termination of a delayed response group sequential test

5.1 Introduction

Suppose we conduct a two treatment comparison, where there is a delay Δ_t in the response of direct clinical interest. Let $X_{A,i} \sim N(\mu_A, \sigma^2)$ and $X_{B,i} \sim N(\mu_B, \sigma^2)$, $i = 1, 2, \ldots$, represent the responses of those subjects allocated to the new treatment and control respectively, where σ^2 is assumed to be known. Our objective is to make inferences about $\theta = \mu_A - \mu_B$, the effect size for our new treatment. In particular, we wish to answer the question "is our new treatment better than control?" In Chapter 2, we introduced a new group sequential design for testing the associated null hypothesis, $H_0 : \theta \leq 0$, which provides a proper treatment of the overrun data that will accumulate should the test stop early. However, we would like a more complete description of the data generated by a trial than a simply stating whether we reject or accept H_0; calculating p-values for testing H_0 and confidence intervals for θ are means of doing just this. In this chapter, we derive such methods of inference for on termination of the GSTs for delayed responses formulated in Chapter 2.

Note that we can think of the one-sided tests of $H_0 : \theta \leq 0$ formulated in Chapter 2 as tests of $H_0 : \theta = 0$ against $H_1 : \theta > 0$ at significance level α. For ease of presentation, in this chapter, we consider inference on termination of tests of this simple null hypothesis. Inference should be based on a sufficient statistic for θ, denoted by $\mathcal{S}$. Let Ω be the sample space for $\mathcal{S}$ defined by a given test of $H_0 : \theta = 0$, i.e., the set of values of $\mathcal{S}$ with which the test can terminate. Upon observing $\mathcal{S} = s^\star$, the p-value for testing

$H_0 : \theta = 0$ is the minimum significance level under which a test defined on Ω can reject H_0 based on an observed outcome $s^\star$; smaller p-values indicate a greater discrepancy between H_0 and the observed data. One natural question is how should one define tests of H_0 on Ω at different significance levels? To resolve this issue, we devise an ordering of the points of Ω which is then implicit in the construction of p-values and confidence intervals. "More extreme" outcomes in this ordering will define the rejection regions of tests of H_0 on Ω at smaller significance levels: points are deemed more extreme if they are higher up the ordering when constructing tests of H_0 against $\theta > 0$, but lower down the ordering for tests of H_0 against $\theta < 0$. Hence, the p-value for testing H_0 upon observing $\mathcal{S} = s^\star$ can be written as

$$\mathbb{P}(\text{Observe } s \text{ as or more extreme than } s^\star; \theta = 0),$$

where the precise definition of what constitutes a more extreme outcome depends on whether we wish to calculate a p-value for testing $H_0 : \theta = 0$ against $\theta > 0$, $\theta < 0$ or $\theta \neq 0$.

Introducing some notation, for $s_1, s_2 \in \Omega$ we write $s_1 \succ s_2$ if s_1 is higher up the ordering than s_2. Furthermore, writing $s_1 \succeq s_2$ indicates that s_1 is higher up or equal to s_2 in the ordering. Let p^+ denote the one-sided upper p-value for testing H_0 against $\theta > 0$ and p^- denote the one-sided lower p-value for testing H_0 against $\theta < 0$. Then, we can write

$$p^+ = \mathbb{P}(\mathcal{S} \succeq s^\star; \theta = 0) \qquad p^- = \mathbb{P}(\mathcal{S} \preceq s^\star; \theta = 0),$$

where $p^+ = 1 - p^-$. The 2-sided p-value for testing H_0 against $\theta \neq 0$ is given by $2 \min\{p^+, p^-\}$.

In order to explain how to find a $(1-\alpha)100\%$ equal-tailed confidence interval for θ, first note that for any choice of θ_0, we can find points in the sample space $s_l(\theta_0)$ and $s_u(\theta_0)$ satisfying

$$\mathbb{P}(\mathcal{S} \succeq s_u(\theta_0); \theta = \theta_0) = \alpha/2$$
$$\mathbb{P}(\mathcal{S} \preceq s_l(\theta_0); \theta = \theta_0) = \alpha/2$$

numerically using a bisection routine. The acceptance set defined by these quantiles, $A(\theta_0) = \{s : s_l(\theta_0) \prec s \prec s_u(\theta_0)\}$, defines the acceptance region of a two-sided test of $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ with type I error rate α. Define

$$Y = \{\theta_0 : \mathcal{S} \in A(\theta_0)\}.$$

We claim Y is a $(1-\alpha)100\%$ equal-tailed confidence set for θ. To prove this, suppose

that the true value of θ is $\tilde{\theta}$. By definition, $\tilde{\theta} \in Y$ if and only if $\mathcal{S} \in A(\tilde{\theta})$. Then,

$$\begin{aligned}\mathbb{P}(\tilde{\theta} \in Y; \theta = \tilde{\theta}) &= \mathbb{P}(\mathcal{S} \in A(\tilde{\theta}); \theta = \tilde{\theta}) \\ &= 1 - \alpha,\end{aligned}$$

as required. If for all s, $\mathbb{P}(\mathcal{S} \succeq s; \theta)$ is an increasing function of θ, we say the distribution of $\mathcal{S}$ is stochastically ordered on Ω with respect to θ and Y will be a $(1-\alpha)100\%$ equal-tailed confidence interval for θ. To illustrate the concepts introduced so far, we now consider how one might order the sample space defined by a fixed sample and group sequential test of $H_0 : \theta = 0$ when response is immediate.

5.1.1 Inference on termination of a fixed sample test

Consider the most simple of tests of $H_0 : \theta = 0$ which stipulates that data are analysed once at the end of the study. Standard theory tells us that the standardised statistic Z based on all accumulated data is a sufficient statistic for θ and we let $\mathcal{S} = Z$. Then, the sample space defined by the fixed sample test is $\Omega_{fix} = \mathbb{R}$. Suppose the test terminates with $Z = z^\star$. Larger values of Z are typical of larger values of θ so that for $\theta_1 > \theta_0$,

$$\frac{f_Z(z_1; \theta_1)}{f_Z(z_1; \theta_0)} > \frac{f_Z(z^\star; \theta_1)}{f_Z(z^\star; \theta_0)} \quad \text{for } z_1 > z^\star,$$

and the monotone likelihood ratio property is said to hold on this sample space. Hence, there is a single natural ordering of Ω_{fix} which stipulates that for $z_1, z_2 \in \Omega_{fix}$, $z_1 \succ z_2$ if $z_1 > z_2$. Then, upon observing $Z = z^\star$, the one-sided upper and lower p-values for testing $H_0 : \theta = 0$ are $p^+ = \mathbb{P}(Z \geq z^\star; \theta = 0)$ and $p^- = \mathbb{P}(Z \leq z^\star; \theta = 0)$, respectively.

5.1.2 Inference on termination of a group sequential test

Suppose we implement our test of $H_0 : \theta = 0$ monitoring the data as each group of responses accumulates. Let Z_k denote the standardised statistic at analysis k. After group $k = 1, \ldots, K-1$, we continue to stage $k+1$ if $l_k < Z_k < u_k$, otherwise we stop and make a decision, either rejecting or accepting H_0. In the absence of early stopping, the test must terminate with a decision at analysis K. One-sided and two-sided tests of H_0 both follow this general form. Our GST generates the sequence of standardised test statistics $\{Z_1, \ldots, Z_K\}$ corresponding to information levels $\{I_1, \ldots, I_K\}$. Define

$$T = \min\{k : Z_k \notin \mathcal{C}_k\},$$

where $\mathcal{C}_k = (l_k, u_k)$ is the continuation region at stage $k = 1, \ldots, K$ and $\mathcal{C}_K = \emptyset$ since in the absence of early stopping, termination must occur at this final analysis. Jennison & Turnbull (2000, Section 8.2) show that the pair (I_T, Z_T) are a sufficient statistic for

θ. Since I_T is indexed by T, we can, equivalently, define the sample space in terms of the pair (T, Z_T). We shall describe inferences in terms of this sufficient statistic. The sample space defined by the GST is

$$\Omega_{\text{GST}} = \{(k, z_k) \,:\, k = 1, \ldots, K \text{ and } z_k \notin \mathcal{C}_k\}. \tag{5.1}$$

We see that the form of Ω_{GST} is more complex than the sample space defined by the fixed sample test, which is simply $\mathbb{R}$. Calculating "naive" fixed sample p-values and confidence intervals on termination of a GST is not appropriate. For example, Tsiatis, Rosner & Mehta (1984) find that the true coverage probabilities of naive 90% confidence intervals calculated on termination of a five-stage two-sided GST of $H_0 : \theta = 0$ of size $\alpha = 0.05$ vary between 88.1% and 93.0%. The attained coverage rate is found to depend on the true value of θ and the shape of the test boundaries.

The monotone likelihood principle does not hold on the sample space defined by a GST because it is possible to terminate at different stages with different information levels. To illustrate this, consider a two-stage GST of $H_0 : \theta = 0$ at equally spaced information levels, with maximum information level $I_2 = 9.85$. Let $L(\theta; k, z_k)$ denote the likelihood function for θ given the test stops with $(T, Z_T) = (k, z_k)$. Then,

$$L(\theta;\, 1,\, z_1) > L(\theta;\, 2, z_2) \qquad \text{if } z_1 > \sqrt{2}\,(z_2 - \theta\sqrt{I_2}/4).$$

Hence, using the likelihood ratio test to test $H_0 : \theta = 0$ versus $H_1 : \theta = 1$, stopping with $(T, Z_T) = (1, 2)$ constitutes more compelling evidence against H_0 than $(T, Z_T) = (2, 2)$. Rosner & Tsiatis (1988) note that since there is no monotone likelihood ratio for testing H_0, there is no uniformly most powerful test according to which we can order points in Ω_{GST}. Hence, it follows that there is no single natural ordering. Several candidates however, have been proposed: the stage-wise, maximum likelihood estimate (mle), likelihood ratio and score statistic ordering. In the likelihood ratio ordering, outcomes are ordered according to the observed value of Z_T, with

$$(k_2, z_2) \succ (k_1, z_1) \qquad \text{if } z_2 > z_1.$$

This ordering is so called because Chang (1989) shows that it is induced by the signed likelihood ratio test on Ω_{GST} of $H_0 : \theta = 0$ versus a general alternative $H_1 : \theta = \theta_1$. In the score statistic ordering, sample points (k, z) are ordered according to the size of the corresponding score statistic for testing H_0, i.e.,

$$(k_2, z_2) \succ (k_1, z_1) \qquad \text{if } z_2\sqrt{I_{k_2}} > z_1\sqrt{I_{k_1}}.$$

Rosner & Tsiatis (1988) note that evaluating $\partial \log L(\theta; k, z_k)/\partial\theta$ at $\theta = 0$ yields the

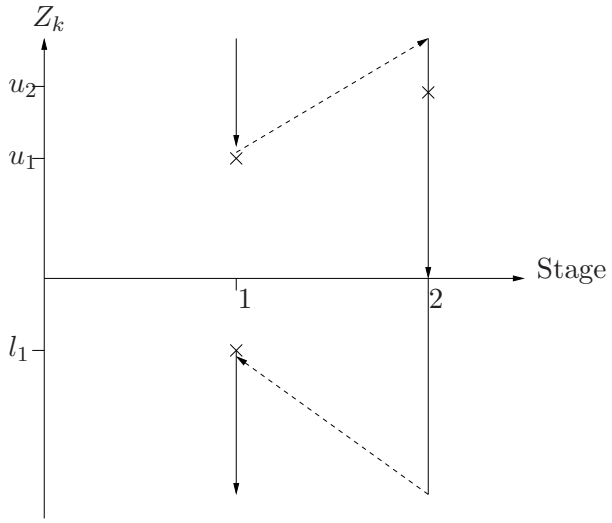


Figure 5-1: Illustrating the stage-wise ordering of the sample space defined by a two-stage one-sided GST of $H_0 : \theta = 0$ against $H_1 : \theta > 0$. Arrows point from outcomes higher in the ordering to those lower in the ordering. As we move down the ordering, the strength of the evidence against H_0 decreases.

statistic $Z_T\sqrt{I_T}$ upon which the score test of $H_0 : \theta = 0$ is based. This score test is locally most powerful for testing $H_0 : \theta = 0$ versus $H_1 : \theta = \theta_1$ for θ_1 close to zero. Hence, ordering outcomes by the size of the score statistic is efficient for testing $\theta = 0$ against local alternatives.

Ordering points in Ω_{GST} according to the mle on termination was first proposed by Armitage (1958) for binomial data and later investigated by Emerson & Fleming (1990) for normal data. Under this ordering,

$$(k_2, z_2) \succ (k_1, z_1) \quad \text{if } z_2/\sqrt{I_{k_2}} > z_1/\sqrt{I_{k_1}} \text{ or equivalently } \hat{\theta}_{k_2} > \hat{\theta}_{k_1}.$$

The stage-wise ordering was first proposed by Armitage in a fully sequential setting and later applied by Fairbanks & Madsen (1982) and Tsiatis et al. (1984) on termination of a GST. As illustrated in Figure 5-1, there are three strata implicit in this ordering. Outcomes (k, z) are first ordered by the boundary crossed, then by stage of termination and finally by the value of z; pairs corresponding to stopping at the first stage with $Z_1 > u_1$ lie at the top of the ordering. More formally, $(k_1, z_1) \succ (k_2, z_2)$ if

$$\begin{aligned} &\text{(i)} \quad k_1 = k_2 \text{ and } z_1 > z_2, \\ &\text{(ii)} \quad k_1 < k_2 \text{ and } z_1 \geq u_{k_1}, \\ &\text{(iii)} \quad k_1 > k_2 \text{ and } z_2 \leq l_{k_2}. \end{aligned} \tag{5.2}$$

Referring to Figure 5-1, one can see that according to this ordering, $(1, u_1) \succ (2, u_2)$, for example, and $(2, u_2) \succ (1, l_1)$.

When choosing between the four orderings described above, there are several desiderata we look for:

1. *P-values agree with the test of H_0.* To ensure internal consistency for the testing procedure, we should observe a p-value $\leq \alpha$ if and only if the GST stops with rejection of H_0.
2. *Monotonicity conditions hold.* The distribution of (T, Z_T) should be stochastically ordered on Ω_{GST} for the chosen ordering with respect to θ. This property ensures the $(1-\alpha)100\%$ confidence set for θ on termination of the GST is an interval.
3. *P-values do not depend on future information levels.* Error spending tests will attain the nominal type I error rate at $\theta = 0$ under any sequence of information levels. Hence, they are often needed in practice when information levels are unpredictable. It will only be possible to compute p-values and confidence intervals on termination in such cases if the chosen ordering of Ω_{GST} does not depend on the number and information levels of future looks at the data.

While each of the four orderings discussed above may have some intuitive appeal, the stage-wise ordering is unique in that it has all three desiderata: p-values based on the other proposed orderings will all depend on future information levels beyond the observed stage of stopping. Jennison & Turnbull (2000, Section 8.4) and Proschan et al. (2006, Section 7.3) both advocate the usage of the stage-wise ordering of the sample space. Desideratum 3 is particularly important since error spending tests are so widely used in practice. Kim & DeMets (1987) give examples of calculating confidence intervals on termination of an error spending test under the stage-wise ordering of the sample space.

So far in this section, we have shown how to make inferences on termination of a standard GST when response is immediate and the flow of data stops upon closure of recruitment. If there is a delay in response however, should the stopping rule be satisfied at an interim analysis and recruitment closed, data will continue to accrue as the responses of pipeline subjects accumulate; we say the test has overrun. Hence, the sample space defined by the test will no longer be of the form (5.1). In the following sections, we discuss some approaches that have been proposed in the literature for dealing with the problem of how one can order the sample space of a GST which has overrun and calculate p-values for testing $H_0 : \theta = 0$.

5.2 P-values after a group sequential test has overrun

5.2.1 Formulation of the problem

Suppose there is a delay of length Δ_t in the endpoint of direct clinical interest. Data are monitored as they accumulate according to a standard group sequential design and recruitment is automatically closed once n_{max} subjects have been accrued. For $k = 1, \ldots, K$, let Z_k denote the standardised statistic and I_k our information for θ at analysis k. At each interim analysis $k = 1, \ldots, K-1$, there will be subjects in the pipeline whose responses have yet to be observed. If the stopping rule is satisfied, i.e., $Z_k \geq u_k$ or $Z_k \leq l_k$, we are obliged to wait for these individuals to respond and report the follow-up. Define $\tilde{Z}_k$ and $\tilde{I}_k$ to be the standardised test statistic and information for θ, respectively, at the re-analysis incorporating the pipeline data. Note that analysis K is scheduled once all n_{max} recruited subjects have been observed. For ease of presentation, we define $\tilde{I}_K = I_K$ and $\tilde{Z}_K = Z_K$. We conclude that the standard GST with analyses scheduled at the information levels $I_1, \ldots, I_K$ will terminate with information level belonging to the set $\{\tilde{I}_1, \ldots, \tilde{I}_K\}$.

Define $T = \min\{k : Z_k \notin \mathcal{C}_k\}$, where $\mathcal{C}_K = \emptyset$. Following the notation introduced in Chapter 2, let $\tilde{Z}_T$ and $\tilde{I}_T$ denote the Z-statistic and information for θ calculated once the flow of data has stopped after the stopping rule has been satisfied at analysis T. In Section 5.8, we prove that $(\tilde{I}_T, \tilde{Z}_T)$ is a sufficient statistic for θ. Once the pipeline data have been observed, a number of things may occur. One hopes that the trend which caused the test statistic to satisfy the stopping rule at the interim analysis persists. A loss of trend however, is also possible. The sample path may have been on a "random high" or "random low" at the interim analysis; the effect of this positive or negative noise can be diluted by the addition of the pipeline data. The GST can terminate with values of $\tilde{Z}_T$ in $\mathbb{R}$ and $\tilde{I}_T$ in the set $\{\tilde{I}_1, \ldots, \tilde{I}_K\}$. Hence, the underlying sample space defined by the overrunning GST, Ω_{O}, is

$$\Omega_{\mathrm{O}} = \cup_{k=1}^{K} \{ (\tilde{I}_k, \tilde{z}_k) : \tilde{z}_k \in \mathbb{R}\}.$$

In contrast to the form of the sample space defined by a standard GST when response is immediate, as given in (5.1), now, when response is delayed, for each k, $\tilde{z}_k$ is allowed to take any real value. Hence, we can drop the subscript k from $\tilde{z}_k$, to obtain

$$\Omega_{\mathrm{O}} = \cup_{k=1}^{K} \{ (\tilde{I}_k, \tilde{z}) : \tilde{z} \in \mathbb{R}\}. \tag{5.3}$$

We adopt this definition of Ω_{O} from now on.

Since there is no single natural ordering on Ω_{O}, the question remains of how one should calculate a p-value for $H_0 : \theta = 0$ when a standard GST has unexpectedly overrun. Several approaches have been proposed in the literature. Hall & Liu (2002) consider both the fully sequential and group sequential cases. They propose methods when the extent of overrunning can be modelling as a function of the stage of termination and p-values are based on a maximum likelihood ordering of the sample space. Hall & Ding (2001) propose an alternative approach based on combining two p-values: one calculated for the sequential part of the test and one for the overrun data assuming these were generated by an independent fixed sample study. In the next section, we explore the deletion method, proposed by Whitehead (1992) and investigated by Sooriyarachchi et al. (2003).

5.2.2 The deletion method

On termination of a GST which has overrun, the sufficient statistic for θ is $(\tilde{I}_T, \tilde{Z}_T)$ rather than (I_T, Z_T). Hence, given that the test terminates at stage $k^\star$ with $\tilde{Z}_{k^\star} = z^\star$, the distribution of $Z_{k^\star}$ does not depend on θ. The deletion method is based on the premise therefore, that no information about θ is lost by deleting from the records the interim analysis at which the stopping rule was first satisfied. To illustrate the method for calculating p-values, suppose the stopping rule is satisfied at analysis $T = k^\star < K$, with either $Z_{k^\star} \geq u_{k^\star}$ or $Z_{k^\star} \leq l_{k^\star}$. We follow-up the pipeline subjects and denote the observed value of $(\tilde{I}_T, \tilde{Z}_T)$ by $(\tilde{I}_{k^\star}, z^\star)$. The deletion method stipulates that interim analysis $k^\star$ be deleted from the records; it is as if the test was originally designed with the first $k^\star$ interim analyses at information levels $I_1, \ldots, I_{k^\star-1}, \tilde{I}_{k^\star}$ and with upper and lower boundary points $(u_1, \ldots, u_{k^\star-1}, z^\star)$ and $(l_1, \ldots, l_{k^\star-1}, z^\star)$. No account is taken in this analysis of possible overrun data if the test had stopped earlier; should we reach analysis K when the test does not overrun, p-values are calculated as if response had in fact been immediate. In effect the deletion method is replacing the analysis which triggered termination of the test with the re-analysis including the pipeline data. Hence, in the spirit of this method, we still refer to the sufficient statistic for θ for this redefined problem as (I_T, Z_T), noting that in our example it takes the value $(\tilde{I}_{k^\star}, z^\star)$.

After stopping at analysis $T = k^\star < K$, the deletion p-value is calculated as if the sample space for (I_T, Z_T) is

$$\Omega_{\mathrm{DEL}} = \cup_{k=1}^{k^\star-1}\{(I_k, z_k) \,:\, z_k \notin \mathcal{C}_k\} \cup \{(\tilde{I}_{k^\star}, z) \,:\, z \in \mathbb{R}\}. \tag{5.4}$$

The definition of Ω_{DEL} depends upon the observed stage of termination; stopping at a different analysis leads to a re-definition of the sample space. For example, suppose we stop at interim analysis $k^\star + 1$ instead of $k^\star$. The deletion p-value is calculated as

if the sample space is

$$\Omega_{\text{DEL}} = \cup_{k=1}^{k^\star}\{(I_k, z_k) \,:\, z_k \notin \mathcal{C}_k\} \cup \{(\tilde{I}_{k^\star+1}, z) \,:\, z \in \mathbb{R}\}. \tag{5.5}$$

Comparing (5.5) with (5.4), we see that stopping with $(\tilde{I}_{k^\star}, z^\star)$ is now not envisaged in this definition of the sample space. We conclude that under the deletion method of calculating p-values, the distribution of (I_T, Z_T) is not properly defined, a fact which has not been pointed out in the literature before. This is an early indication that there will be problems for "p-values" based on this sample space, a point that shall be explored in greater detail in the next sections.

The deletion method proceeds by ordering points of Ω_{DEL} in the spirit of the stage-wise ordering of (5.2). Then, should termination be triggered by exiting an upper boundary, the one-sided upper deletion p-value for testing H_0 upon observing $(\tilde{I}_{k^\star}, z^\star)$ is

$$p^+_{\text{DEL}} = \mathbb{P}\{(I_T < \tilde{I}_{k^\star},\, Z_T \geq u_T) \text{ or } (I_T = \tilde{I}_{k^\star},\, Z_T \geq z^\star)\,;\theta = 0\}.$$

The one-sided lower deletion p-value can then be calculated using the relation $p^-_{\text{DEL}} = 1 - p^+_{\text{DEL}}$. The two-sided p-value for testing $H_0 : \theta = 0$ is given by $p_{\text{DEL}} = 2\min\{p^-_{\text{DEL}}, p^+_{\text{DEL}}\}$. If the GST closes at the first stage with $(I_T, Z_T) = (\tilde{I}_1, z^\star)$ the deletion p-value resolves to the usual fixed sample p-value. In the notation of our delayed response tests, we can write

$$p^+_{\text{DEL}} = \mathbb{P}(\tilde{Z}_1 \geq z^\star; \theta = 0).$$

Should we reach analysis K, the deletion p-value for testing H_0 is the immediate response p-value based on a stage-wise ordering of the sample space $\{(I_k, z_k) \,:\, k = 1, \ldots, K \text{ and } z_k \notin \mathcal{C}_k\}$.

Hall & Liu (2002) suspect that the approach used in practice for the analysis of an overrunning GST will most likely be a variant on the deletion method with different orderings of Ω_{DEL} adopted. Hence, it is of practical interest to investigate properties of the deletion p-value. We have already noted that there are likely to be problems for deletion p-values since the distribution of (T, Z_T) is not well defined. Hall & Liu also point out that the definition of p^+_{DEL} doesn't really reflect the probability of getting to the observed value of the sufficient statistic as it forgets that termination has been triggered precisely because a boundary has been crossed. The authors present empirical evidence to show that in some cases this can result in a marked conservatism in the deletion p-value. In the next sections, we go one step further and prove this result analytically for deletion p-values calculated on termination of two-sided tests of $H_0 : \theta = 0$ with symmetric boundaries and one-sided tests of H_0 with asymmetric

boundaries.

5.3 Properties of deletion p-values

By definition, for continuous responses, a p-value for testing $H_0 : \theta = 0$ should be uniformly distributed on the interval $[0, 1]$ under $\theta = 0$. In particular, under $\theta = 0$, we observe a p-value $\leq \alpha$, for α the significance level of the test, with probability α. In the next section, we prove that two-sided deletion p-values calculated on termination of two-sided tests of H_0 with symmetric boundaries are conservative, i.e., $\mathbb{P}(P_{\text{DEL}} \leq \alpha; \theta = 0) < \alpha$. We then go on to show empirically that this also holds for p^+_{DEL} calculated on termination of one-sided tests of H_0 with asymmetric boundaries.

5.3.1 Deletion p-value on termination of two-sided tests of $H_0 : \theta = 0$

In this section, we consider for the first time in this thesis two-sided GSTs of $H_0 : \theta = 0$. We restrict our attention to two-sided tests of $H_0 : \theta = 0$ with symmetric boundaries which have overrun. We show that on termination of these tests, $\mathbb{P}(P_{\text{DEL}} \leq \alpha; \theta = 0) < \alpha$. In order to do this, we first construct an "exact" p-value for testing H_0 which is uniformly distributed on $[0, 1]$ under H_0. Recall that the sample space for the sufficient statistic $(\tilde{I}_T, \tilde{Z}_T)$ defined by an overrunning GST is

$$\Omega_{\text{O}} = \cup_{k=1}^{K} \{(\tilde{I}_k, \tilde{z}) \, : \, \tilde{z} \in \mathbb{R}\}.$$

When response is immediate and a GST does not overrun, if stopping is triggered by crossing an upper boundary, one usually expects a small one-sided upper p-value; this p-value should be large if the lower boundary is crossed. Hence, when ordering Ω_{O} for the case of delayed responses, we need to split the range of $\tilde{z}$ at each $\tilde{I}_k$ into "high end" and "low end" sections within the overall ordering. We do this by partitioning the sample space about constants h_k, $k = 1, \ldots, K-1$. Sample points are then ordered in the spirit of the stage-wise ordering. We write $(\tilde{I}_{k_1}, \tilde{z}_1) \succ (\tilde{I}_{k_2}, \tilde{z}_2)$ if

$$\begin{aligned} &\text{(i)} \quad \tilde{I}_{k_1} = \tilde{I}_{k_2} \text{ and } \tilde{z}_1 > \tilde{z}_2, \\ &\text{(ii)} \quad \tilde{I}_{k_1} < \tilde{I}_{k_2} \text{ and } \tilde{z}_1 \geq h_{k_1}, \\ &\text{(iii)} \quad \tilde{I}_{k_1} > \tilde{I}_{k_2} \text{ and } \tilde{z}_2 \leq h_{k_2}. \end{aligned} \tag{5.6}$$

For moderate overrun, if termination is triggered by observing $Z_k > u_k$, we expect this to imply $\tilde{Z}_k > h_k$. Likewise, if we observe $Z_k < l_k$, we would expect to observe $\tilde{Z}_k < h_k$ once we incorporate the pipeline responses. However, there may be a small chance of

"switching". We define our p-value for testing H_0 when we observe $(\tilde{I}_{k^\star}, z^\star)$ as

$$\mathbb{P}(\text{Observe } (\tilde{I}_T, \tilde{Z}_T) \text{ as or more extreme than } (\tilde{I}_{k^\star}, z^\star); \theta = 0),$$

where extreme refers to the position of sample points when Ω_{O} is ordered according to (5.6) and partitioned by constants $h_1, \ldots, h_{K-1}$. We denote by p_E^+ and p_E^- our one-sided upper and lower p-values for H_0, respectively. Our two-sided p-value p_E is twice the minimum of these values. It is clear that these p-values will be exact, i.e., have distribution $U(0,1)$ under H_0.

The natural question is how one should choose the partitioning constants $h_1, \ldots, h_{K-1}$. There may be several choices each with some intuitive appeal. However, when the overrun is small, there will usually be a zone of z values with little probability which is some way less than u_k and greater than l_k: h_k can quite happily sit anywhere in this region without making too much difference. This is particularly true at the early stages when boundaries are far apart. The approach we take is to choose the h_k so that we balance switching probabilities, i.e., for each $k = 1, \ldots, K-1$, we choose h_k to satisfy

$$\begin{aligned} &\mathbb{P}(l_1 < Z_1 < u_1, \ldots, l_{k-1} < Z_{k-1} < u_{k-1}, Z_k \geq u_k, \tilde{Z}_k < h_k; \theta = 0) \\ &\quad = \mathbb{P}(l_1 < Z_1 < u_1, \ldots, l_{k-1} < Z_{k-1} < u_{k-1}, Z_k \leq l_k, \tilde{Z}_k \geq h_k; \theta = 0). \end{aligned} \tag{5.7}$$

We refer to (5.7) as our symmetry criterion for choosing partitioning constants. For a two-sided GST of $H_0 : \theta = 0$ with symmetric boundary constants, we obtain $h_1 = \ldots = h_{K-1} = 0$. Figure 5-2 shows how points in the sample space defined by a three-stage symmetric two-sided test of H_0 will be ordered according to (5.6) under this partitioning.

We now consider the form of our p-values under our choice of $h_1 = \ldots = h_{K-1} = 0$. Our two-sided test of H_0 is defined so that at stage k, termination is triggered at stage k if $|Z_k| \geq c_k$. For each $k = 1, \ldots, K-1$, define

$$\begin{aligned} \xi_k(c_1, \ldots, c_k, z^\star) = &\mathbb{P}(|Z_1| < c_1, \ldots, |Z_{k-1}| < c_{k-1}, Z_k \geq c_k, \tilde{Z}_k \geq z^\star; \theta = 0) \\ &+ \mathbb{P}(|Z_1| < c_1, \ldots, |Z_{k-1}| < c_{k-1}, Z_k \leq -c_k, \tilde{Z}_k \geq z^\star; \theta = 0), \end{aligned}$$

and

$$\gamma_k(c_1, \ldots, c_{k-1}, z^\star) = \mathbb{P}(|Z_1| < c_1, \ldots, |Z_{k-1}| < c_{k-1}, Z_k \geq z^\star; \theta = 0).$$

Suppose we stop at stage $T = k^\star < K$ and observe $(\tilde{I}_T, \tilde{Z}_T) = (\tilde{I}_{k^\star}, z^\star)$, with $z^\star \geq h_{k^\star}$.

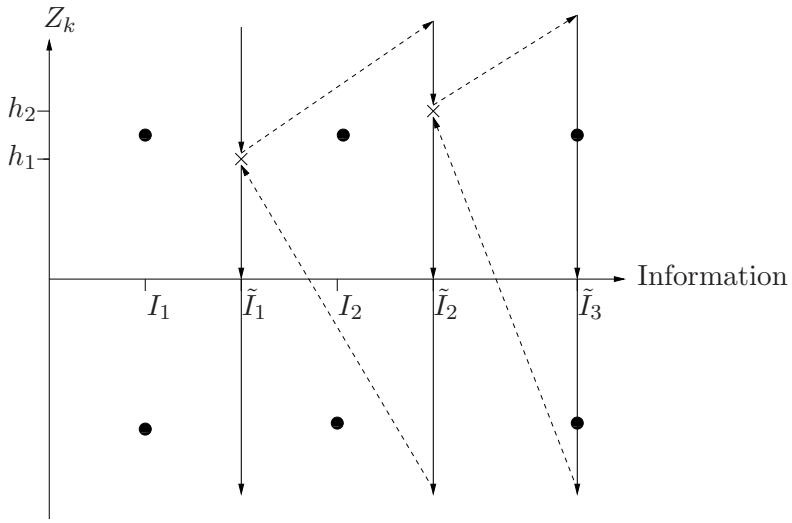


Figure 5-2: Illustration of an ordering of the sample space defined by an overrunning three stage two-sided GST of $H_0 : \theta = 0$. The GST will not overrun if we reach the in the absence of early stopping. If the trial is stopped at an earlier interim analysis it is assumed that the trial will overrun to the same extent.

Then, our one-sided upper p-value for testing H_0 is

$$p_E^+ = \sum_{k=1}^{k^\star-1} \xi_k(c_1, \ldots, c_{k-1}, c_k, 0) + \xi_{k^\star}(c_1, \ldots, c_{k^\star}, z^\star).$$

It follows from the symmetry of our problem that $\gamma_k(c_1, \ldots, c_{k-1}, c_k) = \xi_k(c_1, \ldots, c_k, 0)$, and we can write

$$p_E^+ = \sum_{k=1}^{k^\star-1} \gamma_k(c_1, \ldots, c_{k-1}, c_k) + \xi_{k^\star}(c_1, \ldots, c_{k^\star}, z^\star).$$

If we reach the final stage and observe $(\tilde{I}_K, z^\star)$, then

$$p_E^+ = \sum_{k=1}^{K-1} \gamma_k(c_1, \ldots, c_{k-1}, c_k) + \mathbb{P}(|Z_1| < c_1, \ldots, |Z_{K-1}| < c_{K-1}, \tilde{Z}_K \geq z^\star; \theta = 0).$$

It follows that choosing $h_1, \ldots, h_{K-1}$ according to the symmetry condition (5.7) means that $p_E^+ \leq \alpha/2$ if and only if our test of H_0 terminates either with $z^\star \geq 0$ and $\tilde{I}_{k^\star} < \tilde{I}_K$, or $z^\star \geq c_K$ and $\tilde{I}_{k^\star} = \tilde{I}_K$. This may not be the case under other partitionings of Ω_{O}. The symmetry of our test of problem about $Z = 0$ is also retained. Therefore, from now on we decide to find partitioning constants according to the symmetry condition (5.7) and assume $h_1 = \ldots = h_{K-1} = 0$.

We now compare our exact p-value with the deletion p-value and claim that for any realisation of $(\tilde{I}_T, \tilde{Z}_T)$, $p_{DEL} \geq p_E$, where $p_{DEL} > p_E$ sometimes. Certainly if we reach

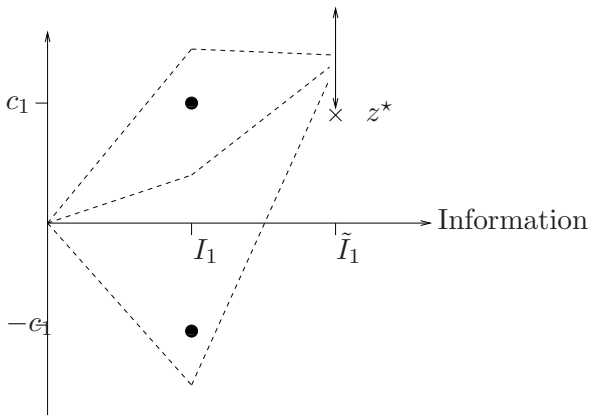


Figure 5-3: Illustration of the form of the sample paths upon which the deletion p-value calculation is based if termination of the trial is triggered at the first interim analysis.

the final stage, we have $p_{DEL} = p_E$. Suppose the test stops at stage $T = k^\star < K$ with $(\tilde{I}_{k^\star}, z^\star)$. First consider the case $z^\star \geq 0$. The symmetry of our problem about zero means the two-sided deletion p-value is $2\mathbb{P}(X_{k^\star}; \theta = 0)$ where

$$\begin{aligned} X_{k^\star} = &\cup_{j=1}^{k^\star-1} \{(z_1, \ldots, z_j, \tilde{z}_j) : z_l \in \mathcal{C}_l, \text{ for } l = 1, \ldots, j-1,\ z_j \geq c_j \text{ and } \tilde{z}_j \in \mathbb{R}\} \\ &\cup \{(z_1, \ldots, z_{k^\star}, \tilde{z}_{k^\star}) : \ z_l \in \mathcal{C}_l, \text{ for } l = 1, \ldots, k^\star - 1,\ z_{k^\star} \in \mathbb{R} \text{ and } \tilde{z}_{k^\star} \geq z^\star\}. \end{aligned}$$

Figure 5-3 depicts elements of the set $X_1 = \{(z_1, \tilde{z}_1) : \ z_1 \in \mathbb{R},\ \tilde{z}_1 \geq z^\star\ \}$. Deleting from the records interim analysis $k^\star$ means that we "forget" that termination was triggered precisely because the sample path had exited the continuation region at $I_{k^\star}$. Our two-sided exact p-value is $2\mathbb{P}(Y_{k^\star}; \theta = 0)$, where

$$\begin{aligned} Y_{k^\star} = &\cup_{j=1}^{k^\star-1} \{(z_1, \ldots, z_j, \tilde{z}_j) : z_l \in \mathcal{C}_l, \text{ for } l = 1, \ldots, j-1,\ z_j \notin \mathcal{C}_j \text{ and } \tilde{z}_j \geq 0\} \\ &\cup \{(z_1, \ldots, z_{k^\star}, \tilde{z}_{k^\star}) : \ z_l \in \mathcal{C}_l, \text{ for } l = 1, \ldots, k^\star - 1,\ z_{k^\star} \notin \mathcal{C}_{k^\star} \text{ and } \tilde{z}_{k^\star} \geq z^\star\}. \end{aligned}$$

The construction of this set acknowledges that the sample path had to fulfill stricter criteria at the interim analysis in order for termination to occur. We have already noted that it follows from our choice of $h_k = 0$ that $\gamma_k(c_1, \ldots, c_{k-1}, c_k) = \xi_k(c_1, \ldots, c_k, 0)$, i.e.,

$$\begin{aligned} &\mathbb{P}(\cup_{j=1}^{k^\star-1} \{(z_1, \ldots, z_j, \tilde{z}_j) : z_l \in \mathcal{C}_l, \text{ for } l = 1, \ldots, j-1,\ z_j \geq c_j \text{ and } \tilde{z}_j \in \mathbb{R}\}; \theta = 0) \\ &\quad = \mathbb{P}(\cup_{j=1}^{k^\star-1} \{(z_1, \ldots, z_j, \tilde{z}_j) : z_l \in \mathcal{C}_l, \text{ for } l = 1, \ldots, j-1,\ z_j \notin \mathcal{C}_j \text{ and } \tilde{z}_j \geq 0\}; \theta = 0). \end{aligned}$$

Then, comparing $X_{k^\star}$ with $Y_{k^\star}$ one can deduce that $p_{\text{DEL}} > p_E$.

Now consider the case where $z^\star < 0$. The two-sided deletion p-value is given by

$p_{\mathrm{DEL}} = 2\mathbb{P}(X'_{k^\star}; \theta = 0)$, where

$$X'_{k^\star} = \cup_{j=1}^{k^\star - 1} \{(z_1, \ldots, z_j, \tilde{z}_j) : z_l \in \mathcal{C}_l, \text{ for } l = 1, \ldots, j-1,\ z_j \leq l_j \text{ and } \tilde{z}_j \in \mathbb{R}\}$$
$$\cup \{(z_1, \ldots, z_{k^\star}, \tilde{z}_{k^\star}) : \ z_l \in \mathcal{C}_l, \text{ for } l = 1, \ldots, k^\star - 1,\ z_{k^\star} \in \mathbb{R} \text{ and } \tilde{z}_{k^\star} \leq z^\star\}.$$

Meanwhile, our exact two-sided p-value is $2\mathbb{P}(Y'_{k^\star}; \theta = 0)$, where

$$Y'_{k^\star} = \cup_{j=1}^{k^\star} \{(z_1, \ldots, z_j, \tilde{z}_j) : z_l \in \mathcal{C}_l, \text{ for } l = 1, \ldots, j-1,\ z_j \notin \mathcal{C}_j \text{ and } \tilde{z}_j < 0\}$$
$$\cup \{(z_1, \ldots, z_{k^\star}, \tilde{z}_{k^\star}) : \ z_l \in \mathcal{C}_l, \text{ for } l = 1, \ldots, k^\star - 1,\ z_{k^\star} \notin \mathcal{C}_{k^\star} \text{ and } \tilde{z}_{k^\star} \leq z^\star\}.$$

Again, it follows from our choice of $h_k = 0$ to satisfy criterion (5.7) that $\mathbb{P}(X'_{k^\star}; \theta = 0) > \mathbb{P}(Y'_{k^\star}; \theta = 0)$ and $p_{\mathrm{DEL}} > p_E$. Our claim that $p_{\mathrm{DEL}} \geq p_E$ for all realisations of $(\tilde{I}_T, \tilde{Z}_T)$, where this inequality is strict for $\tilde{I}_{k^\star} < \tilde{I}_K$, is proved. Since $\mathbb{P}(P_E \leq \alpha; \theta = 0) = \alpha$ it follows that the deletion p-value is indeed conservative.

5.3.2 An example

We explore the conservatism of the deletion p-value using a simple example. Sooriyarachchi et al. (2003) investigate via simulation the accuracy of the deletion p-value calculated on termination of an O'Brien & Fleming (1979) test (OBF). They consider a two-treatment comparison trial in stroke. We deviate from their example slightly and assume we observe independent responses $X_{A,i} \sim N(\mu_A, \sigma^2)$ and $X_{B,i} \sim N(\mu_B, \sigma^2)$, $i = 1, 2, \ldots$, on the new and control treatments, respectively. Define $\theta = \mu_A - \mu_B$. Data are to be analysed in $K = 5$ equally sized groups, with equal numbers on each treatment. At each analysis, we apply the stopping rule of a $K = 5$-stage OBF test of $H_0 : \theta = 0$ against $H_1 : \theta \neq 0$. The test is designed to attain type I error rate $\alpha = 0.05$ at $\theta = 0$ and power $1 - \beta = 0.9$ at $\theta = \pm 0.56$. We assume σ^2 is known and such that the OBF test requires a maximum sample size of $n_{max} = 450$ subjects, i.e., responses are to be analysed in groups of 90. Accrual proceeds at 15 subjects per month. The endpoint of clinical interest is response at 90 days after stroke and an extra one month delay in data transfer was anticipated; in our notation, it is as if $\Delta_t = 4$ months and data transfer is immediate. Hence, at each interim analysis, 60 subjects will be in the pipeline which we are obliged to follow-up should termination be triggered. At the stated rate of accrual, recruitment will be completed after $t_{max} = 450/15 = 30$ months and $r = \Delta_t / t_{max} = 2/15$. Reading from Table 5.1, we see that for this example $\mathbb{P}(P_{\mathrm{DEL}} \leq \alpha; \theta = 0) = 0.046$ and the deletion p-value is conservative as expected.

Table 5.1 lists values of $\mathbb{P}(P_{\mathrm{DEL}} \leq \alpha; \theta = 0)$ for Pocock and OBF tests designed to attain type I error rate α at $\theta = 0$ and power $1 - \beta$ at $\theta = \delta$ which have unexpectedly

r	Pocock test	OBF test
0.01	0.050	0.050
0.1	0.038	0.047
0.13	0.035	0.046
0.2	0.030	0.044
0.3	0.026	0.043
0.4	0.024	0.043

Table 5.1: Probabilities of observing $P_{\text{DEL}} \leq \alpha$ under $\theta = 0$ on termination of Pocock and OBF tests of $H_0 : \theta = 0$. All tests have $K = 5$, $\alpha = 0.05$ and are designed to attain power $1 - \beta = 0.9$ at $\theta = \pm\delta$.

overrun. Data are analysed in $K = 5$ equally sized groups of observations, with equal numbers on each treatment. Results are invariant to changes in δ and σ^2 if r remains equal to the stated value. The conservatism of the deletion method depends on the shape of the stopping boundaries and performs worse for the Pocock test. Indeed, for $r = 0.4$, we are in fact testing at below significance level $\alpha/2$ using the deletion method. The difference in performance of the p-value can be explained by consideration of the way the Pocock and OBF tests spend their type I error probabilities. The OBF spends most of α at the final analysis when there is no overrunning data and the deletion p-value coincides with our exact p-value. However, testing is more aggressive under the Pocock test; there is a greater probability that we stop at an early interim analysis when the test will overrun and the deletion p-value will be conservative.

5.3.3 Properties of one-sided deletion p-value

Suppose we test $H_0 : \theta = 0$ against $H_1 : \theta > 0$ using a $K = 5$-stage design from the power family of one-sided tests of Pampallona & Tsiatis (1994) with shape parameter $\Delta = -0.5$. Figure 5-4 shows that the one-sided upper deletion p-value for testing H_0 is again conservative when $\theta = 0$ at level α, although probabilities remain within 0.001 of their desired values for values of $r < 0.5$. This conservatism is bourne out in Figure 5-5, which plots the cumulative distribution function of P^+_{DEL} under $\theta = 0$ when $r = 0.4$. Its deviation from the line $\mathbb{P}(P^+_{DEL} \leq p; \theta = 0) = p$ shows that P^+_{DEL} does not have $U(0, 1)$ distribution when $\theta = 0$ and so is not a proper p-value.

So far, we have described methods of inference which can be used to "rescue" the inference of a GST which has unexpectedly overrun. However, we have already formulated efficient designs which provide a proper treatment of the overrun data should we stop at an interim analysis. In the following sections, we shall describe how we can derive exact p-values and $(1 - \alpha)100\%$ equal-tailed confidence intervals on termination of our GSTs designed specifically for delayed responses.

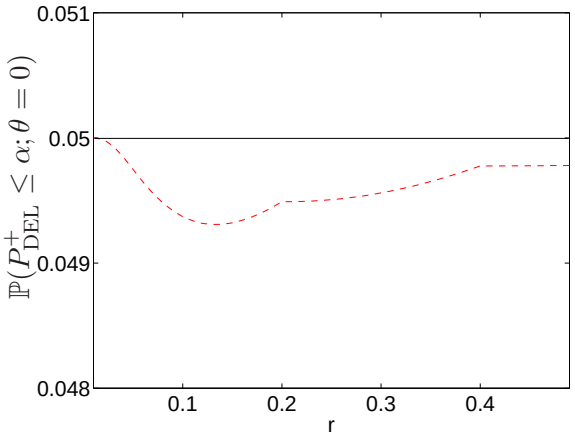


Figure 5-4: Plot of $\mathbb{P}(P^+_{\mathrm{DEL}} \leq \alpha; \theta = 0)$ when the deletion p-value is calculated on termination of a one-sided power test of $H_0 : \theta = 0$ against $\theta > 0$ with shape parameter $\Delta = -0.5$. The test is designed to attain type I error probability $\alpha = 0.05$ at $\theta = 0$, power $1 - \beta = 0.9$ at $\theta = \delta$ with $K = 5$.

5.4 Exact p-values on termination of a delayed response group sequential test

In this section, we describe how one can calculate p-values on termination of a delayed response GST of $H_0 : \theta = 0$ against $H_1 : \theta > 0$ as formulated in Chapter 2. First, we recap on how a K-stage test of this form would proceed. At interim analysis k, where $k \in \{1, \ldots, K-1\}$, termination is triggered if $Z_k \geq u_k$ or $Z_k \leq l_k$; recruitment is closed and we wait for responses in the pipeline before conducting decision analysis k, rejecting H_0 if $\tilde{Z}_k \geq c_k$. At interim analysis $K-1$, if $l_{K-1} < Z_{K-1} < u_{K-1}$, accrual continues until n_{max} subjects are recruited. We then wait until all n_{max} responses are available before conducting decision analysis K, rejecting H_0 if $\tilde{Z}_K \geq c_K$. We denote the information for θ at decision analysis k by $\tilde{I}_k$. Define $T = \min\{k : Z_k \notin \mathcal{C}_k\}$, where $\mathcal{C}_k$ is the continuation region at interim analysis k and we define $\mathcal{C}_K = \emptyset$. In Section 5.8, we prove $(\tilde{I}_T, \tilde{Z}_T)$ is a sufficient statistic on termination for θ. The sample space defined by our test is

$$\Omega = \cup_{k=1}^{K} \{(\tilde{I}_k, z) : z \in \mathbb{R}\}.$$

To formulate an ordering for Ω, we take ideas from Section 5.3.1 used to devise an ordering of the sample space defined by a standard GST which has unexpectedly overrun. For each $k = 1, \ldots, K-1$, we partition the sample space at decision analysis k about the decision constant c_k. Outcomes are then ordered in the spirit of the

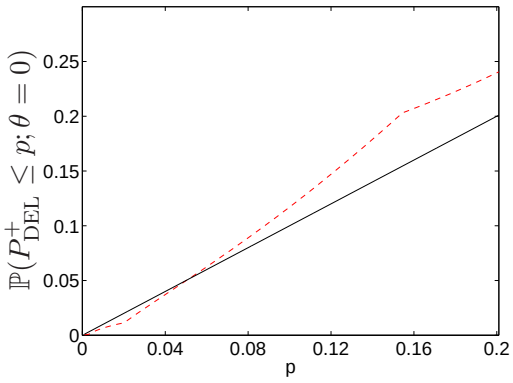


Figure 5-5: Plot of the cumulative distribution function of P^+_{DEL} under $\theta = 0$ when $r = 0.4$. P-values are calculated on termination of a one-sided power test of $H_0 : \theta = 0$ against $\theta > 0$ with shape parameter $\Delta = -0.5$. The test is designed to attain type I error probability $\alpha = 0.05$ at $\theta = 0$, power $1 - \beta = 0.9$ at $\theta = \delta$ with $K = 5$.

stage-wise ordering (5.2). We define $(\tilde{I}_{k_1}, z_1) \succ (\tilde{I}_{k_2}, z_2)$ if

$$\begin{aligned} &\text{(i)} \quad \tilde{I}_{k_1} = \tilde{I}_{k_2} \text{ and } z_1 > z_2, \\ &\text{(ii)} \quad \tilde{I}_{k_1} < \tilde{I}_{k_2} \text{ and } z_1 \geq c_{k_1}, \\ &\text{(iii)} \quad \tilde{I}_{k_1} > \tilde{I}_{k_2} \text{ and } z_2 < c_{k_2}. \end{aligned} \tag{5.8}$$

Figure 5-6 illustrates this ordering on the sample space defined by a three-stage delayed response GST. Let p-values for testing $H_0 : \theta = 0$ against $\theta > 0$ based on the above ordering of Ω be realisations of the random variable P_E. It is clear that $P_E \sim U(0, 1)$ under $\theta = 0$. Based on the above ordering, we obtain a p-value of less than or equal to α if and only if the delayed response GST stops to reject H_0.

Under this ordering, early stopping for rejection of H_0 is associated with larger values of θ. When r is small, this is natural. Figure 5-7 illustrates that sample paths are likely to terminate in a local neighbourhood of their position at the interim analysis. Hence, if stopping is prompted by crossing an upper boundary we expect to observe $\tilde{Z}_k \geq c_k$ at the decision analysis. We conclude therefore, that our proposed ordering of Ω is a natural extension of the stage-wise ordering for the sample space of a GST when response is immediate. Figure 5-8 illustrates that for larger values of r however, crossing an upper boundary does not necessarily imply we will reject H_0 at the decision analysis. The probability of switching, i.e., observing $Z_k \geq u_k$ and then $\tilde{Z}_k < c_k$ or $Z_k \leq l_k$ and then $\tilde{Z}_k \geq c_k$, is now higher. Hence, it may not seem so natural to always associate stopping early for rejection of H_0 with evidence of larger values of θ. However, we do need to do something to combine outcomes with different stages of stopping. This issue could be resolved by ordering Ω according to the value of the mle $\tilde{Z}_T/\sqrt{\tilde{I}_T}$, although

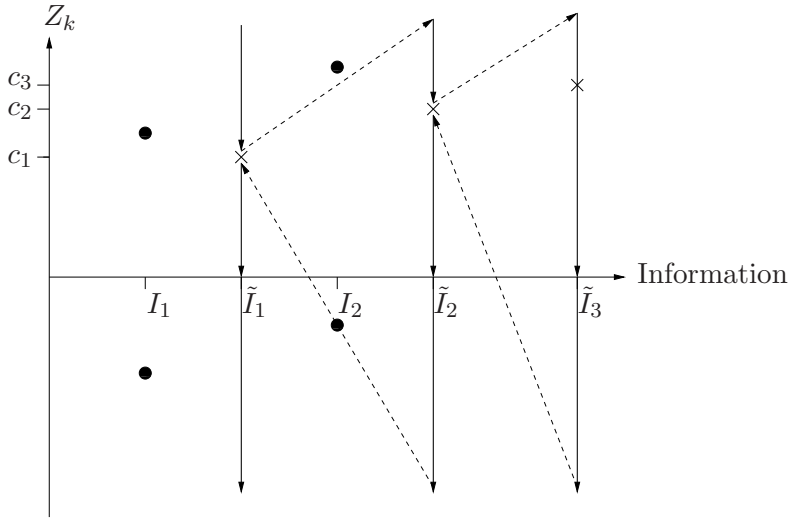


Figure 5-6: Illustration of the stage-wise ordering of the sample space defined by a three-stage delayed response GST. Arrows point from sample points higher in the proposed ordering to those lower in the ordering. Dots mark the boundary values at each interim analysis.

p-values based on this ordering will depend on observed information levels beyond the observed stopping stage $T = k^\star$. If these cannot be predicted exactly and, for example, when using error spending tests, then inference on termination will be not possible. In view of this concern, we adopt the ordering of (5.8) which orders outcomes based on different numbers of observations in the spirit of stage-wise ordering.

5.5 Stochastic ordering of the distribution of $(\tilde{I}_T, \tilde{Z}_T)$ on Ω

A desideratum of an ordering of Ω is that the distribution of $(\tilde{I}_T, \tilde{Z}_T)$ on this sample space be stochastically ordered with respect to θ, i.e., for each $(\tilde{I}_k, z) \in \Omega$, $\mathbb{P}((\tilde{I}_T, \tilde{Z}_T; \theta) \succeq (\tilde{I}_k, z))$ is increasing in θ. We refer to this as the monotonicity property and it ensures that the equal-tailed $(1-\alpha)100\%$ confidence set for θ is an interval. In general, this property will not hold for our proposed ordering on Ω. For example, consider a two-stage delayed response GST which allows stopping at the first stage for futility only, i.e., we set $u_1 = +\infty$. Then,

$$\mathbb{P}((\tilde{I}_T, \tilde{Z}_T; \theta) \succeq (1, c_1)) = \mathbb{P}(Z_1 \leq l_1, \tilde{Z}_1 \geq c_1; \theta), \tag{5.9}$$

where the RHS of (5.9) is decreasing in θ when θ is large. If $u_1 < +\infty$ and $\mathbb{P}(Z_1 \geq u_1, \tilde{Z}_1 \geq c_1; \theta)$ increases with θ, things are not so clear cut. We claim that monotonicity does hold approximately in most cases. In this section, we give bounds for any deviations from monotonicity that may occur on the sample space defined by

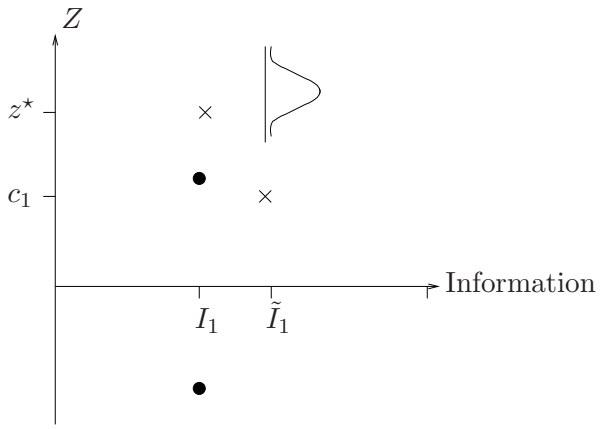


Figure 5-7: Illustration of the likely value of $\tilde{Z}_1$ when termination is triggered at the first interim analysis by observing $Z_1 = z^\star > u_1$ and the delay parameter r is small.

an optimal GST when r is small and for larger values of r, we give empirical evidence that monotonicity still holds.

Consider outcomes in Ω for which a K-stage test of $H_0 : \theta \leq 0$ against $H_1 : \theta > 0$ stops with rejection of H_0. For $k = 1, \ldots, K-1$, define

$$N_k(z^\star) = \{Z_1 \in \mathcal{C}_1, \ldots, Z_{k-1} \in \mathcal{C}_{j-1}, Z_k \geq u_k, \tilde{Z}_k \geq z^\star\}$$
$$M_k(z^\star) = \{Z_1 \in \mathcal{C}_1, \ldots, Z_{k-1} \in \mathcal{C}_{k-1}, Z_k \leq l_k, \tilde{Z}_k \geq z^\star\}.$$

Then, define

$$A_k(z^\star) = \bigcup_{j=1}^{k-1} N_j(c_j) \cup N_k(z^\star), \qquad B_k(z^\star) = \bigcup_{j=1}^{k-1} M_j(c_j) \cup M_k(z^\star),$$

so that the event $\{(\tilde{I}_T, \tilde{Z}_T) \succeq (\tilde{I}_k, z^\star)\} = A_k(z^\star) \cup B_k(z^\star)$. Define

$$A_K(z^\star) = A_{K-1}(c_{K-1}) \cup \{Z_1 \in \mathcal{C}_1, \ldots, Z_{K-1} \in \mathcal{C}_{K-1}, \tilde{Z}_K \geq z^\star\}.$$

We can write the event $\{(\tilde{I}_T, \tilde{Z}_T) \succeq (\tilde{I}_K, z^\star)\}$ as the union of $A_K(z^\star)$ with $B_{K-1}(c_{K-1})$.

We first prove that $\mathbb{P}(A_k(z^\star); \theta)$ is increasing in θ, for $k = 1, \ldots, K$. This can be done using a coupling argument. To see this, suppose $\theta' = \theta + \delta$, where $\delta > 0$. Let $I_{(i)}$ denote the ith smallest information level in the sequence $\{I_1, \tilde{I}_1, \ldots, I_{K-1}, \tilde{I}_{K-1}, \tilde{I}_K\}$ and let $Z_{(i)}$ denote the associated standardised test statistic. Define $\Delta_i = I_{(i)} - I_{(i-1)}$, for $i = 2, \ldots, 2K-1$. Suppose we have independent random variables $Y_1 \sim N(\theta I_1, I_1)$ and $Y_i \sim N(\theta\Delta_i, \Delta_i)$, $i = 2, \ldots, 2K-1$. Under θ, the sequence of standardised test

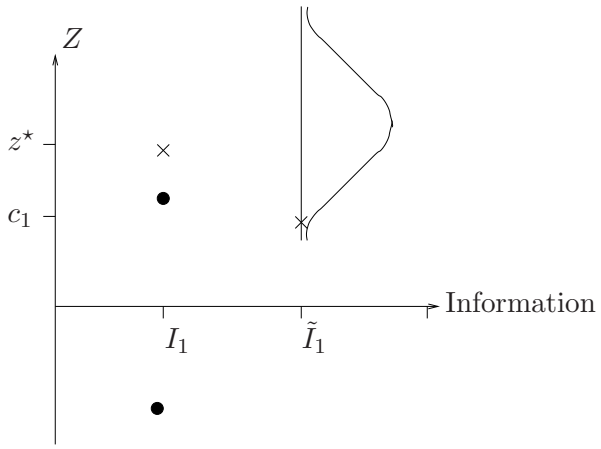


Figure 5-8: Illustration of the likely value of $\tilde{Z}_1$ when termination is triggered at the first interim analysis by observing $Z_1 = z^\star > u_1$ and the delay parameter r is large.

statistics $\{Z_{(1)}, \ldots, Z_{(2K-1)}\}$ has the same joint distribution as

$$\bigcup_{k=1}^{2K-1} \left\{ \frac{1}{\sqrt{I_{(k)}}} \sum_{i=1}^{k} Y_i \right\}.$$

Let $Y_1' = Y_1 + I_1\delta$ and $Y_i' = Y_i + \Delta_i\delta$, $i = 2, \ldots, 2K-1$. We define $Z'_{(1)} = Y_1'/\sqrt{I_1}$ and $Z'_{(i)} = (Y_1' + \ldots + Y_i')/\sqrt{I_{(i)}}$. The sequence $\{Z'_{(1)}, \ldots, Z'_{(2K-1)}\}$ has the appropriate joint distribution under $\theta' = \theta + \delta$. By construction, $Z'_{(i)} > Z_{(i)}$, and it follows that for each $k = 1, \ldots, K$,

$$\mathbb{P}(\, A_k(z^\star); \theta' \,) > \mathbb{P}(\, A_k(z^\star); \theta \,). \tag{5.10}$$

We now find a bound for any loss of monotonicity for the probability of events $\{(\tilde{I}_T, \tilde{Z}_T) \succeq (\tilde{I}_k, z^\star)\}$ with $\tilde{I}_k < \tilde{I}_K$ and $z^\star \geq c_k$. For monotonicity to hold, $\mathbb{P}(A_k(z^\star) \cup B_k(z^\star); \theta)$ must be increasing in θ. Let $g_j(\tilde{z}_j; \theta)$ be the marginal density of $\tilde{Z}_j$ at $\tilde{z}_j$. For each $j = 1, \ldots, k$, for $z^\star \geq c_j$,

$$\begin{aligned}
\mathbb{P}(M_j(z^\star); \theta) &= \int_{z^\star}^{\infty} g_j(\tilde{z}_j; \theta) \mathbb{P}(Z_1 \in \mathcal{C}_1, \ldots, Z_{j-1} \in \mathcal{C}_{j-1}, Z_j \leq l_j | \tilde{z}_j) d\tilde{z}_j \\
&\leq \int_{z^\star}^{\infty} g_j(\tilde{z}_j; \theta) \mathbb{P}(Z_1 \in \mathcal{C}_1, \ldots, Z_{j-1} \in \mathcal{C}_{j-1}, Z_j \leq l_j | c_j) d\tilde{z}_j \\
&< \eta_j,
\end{aligned} \tag{5.11}$$

where $\eta_j = \mathbb{P}(Z_1 \in \mathcal{C}_1, \ldots, Z_{j-1} \in \mathcal{C}_{j-1}, Z_j \leq l_j | c_j)$. If $S_j = Z_j\sqrt{I_j}$, $j = 1, \ldots, k$, and $\tilde{S}_k = \tilde{Z}_k\sqrt{\tilde{I}_k}$, marginally, $\{S_1, \ldots, S_k, \tilde{S}_k\}$ given $\tilde{Z}_k = c_k$ is distributed as a Brownian bridge observed at times $\{I_1, \ldots, I_k, \tilde{I}_k\}$. Hence η_j does not depend on θ. Noting that

r	η_1
0.01	1.82×10^{-16}
0.03	6.10×10^{-7}
0.05	5.31×10^{-5}
0.07	3.64×10^{-4}
0.1	1.53×10^{-3}

Table 5.2: Values of $\eta_1 = \mathbb{P}(Z_1 \leq l_1 | \tilde{Z}_1 = c_1)$ calculated for optimal delayed response GSTs minimising objective function F_2. Tests have $K = 5$, $\alpha = 0.05$, $\beta = 0.1$ and $R = 1.15$ and analyses are scheduled following the pattern (2.3).

$\mathbb{P}(A_k(z^\star); \theta)$ is increasing in θ, we deduce

$$\begin{aligned}
&\mathbb{P}(\, A_k(z^\star) \cup B_k(z^\star); \theta' \,) - \mathbb{P}(\, A_k(z^\star) \cup B_k(z^\star); \theta \,) \\
&\qquad \geq \mathbb{P}(\, B_k(z^\star); \theta' \,) - \mathbb{P}(\, B_k(z^\star); \theta \,) && \text{by (5.10)} \\
&\qquad \geq 0 - \mathbb{P}(B_k(z^\star); \theta) \\
&\qquad = -\sum_{j=1}^{k-1} \mathbb{P}(M_j(c_j); \theta) - \mathbb{P}(M_k(z^\star); \theta) \\
&\qquad > -\sum_{j=1}^{k} \eta_j && \text{for all } \theta' > \theta \text{ by (5.11)}
\end{aligned}$$

Hence, we have a bound on any deviation from monotonicity that can occur for outcomes for which our test terminates early with rejection of H_0. Table 5.2 lists values of η_1 for a 5-stage optimal test of H_0 minimising F_2; for other values of j, η_j is zero to more than 10 decimal places. For small values of $r \leq 0.1$, the bounds on any deviation from monotonicity are very small. For larger values of r, we have studied the case when $K = 2$, $r = 0.4$ and $\eta_1 = 0.176$. This bound is too high for the preceding argument to convince us that we are near monotonicity. However, Figure 5-9 plots $\mathbb{P}((\tilde{I}_T, \tilde{Z}_T) \succeq (\tilde{I}_1, c_1); \theta)$ and $\mathbb{P}((\tilde{I}_T, \tilde{Z}_T) \succeq (\tilde{I}_2, c_2); \theta)$ for this test and shows that the overall probabilities of interest are still increasing in θ. We have also considered other outcomes in Ω for this test and found $\mathbb{P}((\tilde{I}_T, \tilde{Z}_T) \succeq (\tilde{I}_k, z); \theta)$ to be increasing in θ.

When constructing our GSTs of $H_0 : \theta \leq 0$ for delayed responses, we have implicitly assumed that controlling the type I error rate at level α at $\theta = 0$ will ensure that our test is of size α. Under our proposed ordering of Ω, this is equivalent to assuming $\mathbb{P}((\tilde{I}_T, \tilde{Z}_T) \succeq (\tilde{I}_K, c_K); \theta)$ is increasing in θ for null values of the parameter. The checks made in this section to check for monotonicity show that we can have confidence that this is true, even for tests designed under large values of r.

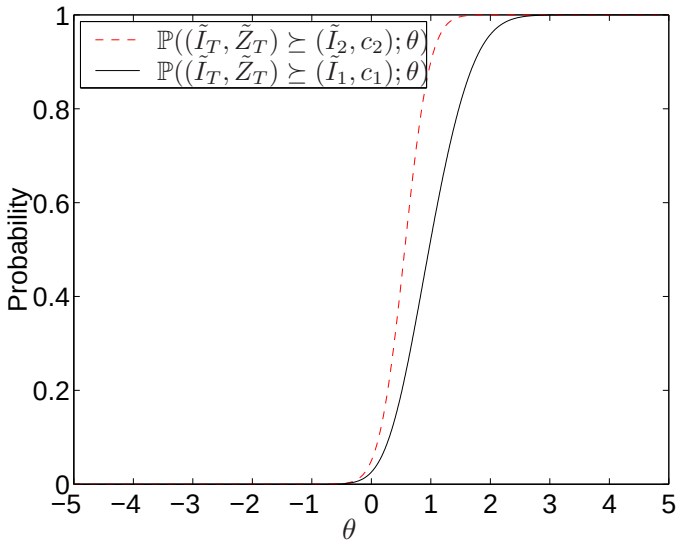


Figure 5-9: Plot of $\mathbb{P}((\tilde{I}_T, \tilde{Z}_T) \succeq (\tilde{I}_1, c_1); \theta)$ and $\mathbb{P}((\tilde{I}_T, \tilde{Z}_T) \succeq (\tilde{I}_2, c_2); \theta)$ for a two-stage test minimising F_2 when $r = 0.4$. The test has $\alpha = 0.05$, $\beta = 0.1$ and $R = 1.15$.

5.6 Confidence intervals on termination of a delayed response group sequential test

Suppose the observed value of $(\tilde{I}_T, \tilde{Z}_T)$ is $(I^\star, z^\star)$. For any θ_0 we can find $(I_l(\theta_0), z_l(\theta_0))$ and $(I_u(\theta_0), z_u(\theta_0))$ such that

$$\mathbb{P}((\tilde{I}_T, \tilde{Z}_T) \preceq (I_u(\theta_0), z_u(\theta_0)); \theta_0) = 1 - \alpha/2$$
$$\mathbb{P}((\tilde{I}_T, \tilde{Z}_T) \preceq (I_l(\theta_0), z_l(\theta_0)); \theta_0) = \alpha/2.$$

Define

$$A(\theta_0) = \{\ (I, z) : (I_l(\theta_0), z_l(\theta_0)) \preceq (I, z) \preceq (I_u(\theta_0), z_u(\theta_0))\ \}.$$

From standard arguments presented in Section 5.1.2, it follows that the set

$$\{\ \theta : (\tilde{I}_T, \tilde{Z}_T) \in A(\theta)\ \}$$

is a $(1 - \alpha)$-level equal tailed confidence set for θ. We can also write this set in the form

$$\{\theta : \mathbb{P}((\tilde{I}_T, \tilde{Z}_T) \preceq (I^\star, z^\star); \theta) \in (\alpha/2, 1 - \alpha/2)\ \}. \tag{5.12}$$

Assuming the distribution of $(\tilde{I}_T, \tilde{Z}_T)$ is stochastically ordered on Ω with respect to θ given the arguments of Section 5.5 then the set (5.12) will be an interval (θ_L, θ_U), where

$$\mathbb{P}((\tilde{I}_T, \tilde{Z}_T) \succeq (I^\star, z^\star); \theta_L) = \mathbb{P}((\tilde{I}_T, \tilde{Z}_T) \preceq (I^\star, z^\star); \theta_U) = \alpha/2.$$

These two equations can be solved for θ_U and θ_L using a bisection search.

As a failsafe, one can check the values of $\mathbb{P}((\tilde{I}_T, \tilde{Z}_T) \succeq (I^\star, z^\star); \theta)$ on a grid of θ values to confirm the confidence set is an interval. To illustrate how one might to this, suppose in practice it is required that the limits of the confidence interval will be rounded to one decimal place. Hence, we lose no accuracy by positioning our grid points at intervals of 0.1, e.g., at 0.05, 0.15, etc.; if results are required to a higher degree of accuracy, the spacing of the grid points can be adjusted accordingly. One then proceeds by moving from the LHS of the grid upwards through the θ values calculating $\mathbb{P}((\tilde{I}_T, \tilde{Z}_T) \preceq (I^\star, z^\star); \theta)$. One stops at the first grid value, $\theta_{i_1^\star}$, for which

$$\mathbb{P}((\tilde{I}_T, \tilde{Z}_T) \preceq (I^\star, z^\star); \theta_{i_1^\star}) < 1 - \alpha/2,$$

and we set $\theta_L = \theta_{i_1^\star - 1} + 0.05$. We find θ_U using a similar approach, by moving from the RHS of the grid downwards through the θ values. We stop at the first grid point, $\theta_{i_2^\star}$, for which

$$\mathbb{P}((\tilde{I}_T, \tilde{Z}_T) \preceq (I^\star, z^\star); \theta_{i_2^\star}) < \alpha/2$$

and set $\theta_U = \theta_{i_2^\star + 1} + 0.05$. If we find that the confidence set is not an interval, one pragmatic solution is to take θ_L and θ_U to be the infimum and supremum of the confidence set respectively. The resulting confidence interval will be conservative in the sense that its coverage rate will exceed $1 - \alpha$.

5.7 Practical implications

5.7.1 Adapting standard group sequential tests for delayed responses

Recall the trial scenario outlined in Section 5.2.1 where it is known at the design stage that a delay is inherent in the primary endpoint. Suppose we prefer not to use one of the optimal delayed response GSTs proposed in Chapter 3; we still want to plan ahead for the overrun data but we want to apply the stopping boundaries of a standard GST at the interim analyses. After all, software for finding optimal tests may not be readily available. In this section, drawing on ideas from Section 5.3.1, we explain how one can adapt standard GSTs for delayed responses to produce tests which are efficient for small values of r.

Suppose we have a K-stage standard GST of $H_0 : \theta \leq 0$ against $H_1 : \theta > 0$ designed to attain type I error probability α at $\theta = 0$ and power $1-\beta$ at $\theta = \delta$ under the information sequence $\{I_1, \ldots, I_K\}$. For each $k = 1, \ldots, K-1$, at interim analysis k, the test has continuation region $\mathcal{C}_k = (l_k, u_k)$. At analysis K, it terminates with rejection of H_0

if $Z_K \geq u_K$ and acceptance otherwise. At interim analysis k, recruitment is closed if $Z_k \geq u_k$ or $Z_k \leq l_k$; we are obliged to wait and follow-up any subjects in the pipeline before re-analysing the data at what we call decision analysis k. We want to plan ahead and choose constants $c_1, \ldots, c_{K-1}$ so that we reject H_0 if $\tilde{Z}_k \geq c_k$ and accept it otherwise. However, it is not immediately clear how one should choose these constants and indeed what the error probabilities of our new test for delayed responses will be. Drawing on the ideas of Section 5.3.1 for the partitioning of the sample space of an overrunning GST, we can choose $c_1, \ldots, c_{K-1}$ so that our new test has type I error rate α at $\theta = 0$, as required.

For each $k = 1, \ldots, K-1$, we choose c_k to satisfy

$$\begin{aligned}\mathbb{P}(Z_1 \in \mathcal{C}_1, \ldots, Z_{k-1} \in \mathcal{C}_{k-1}, Z_k \geq u_k, \tilde{Z}_k < c_k; \theta = 0) \\ = \mathbb{P}(Z_1 \in \mathcal{C}_1, \ldots, Z_{k-1} \in \mathcal{C}_{k-1}, Z_k \leq l_k, \tilde{Z}_k \geq c_k; \theta = 0).\end{aligned}$$

This symmetry condition states that the reversal probabilities under $\theta = 0$ are equal. Define

$$\begin{aligned}\psi_k(l_1, \ldots, l_k, u_k, c_k; \theta) &= \mathbb{P}(Z_1 \in \mathcal{C}_1, \ldots, Z_{k-1} \in \mathcal{C}_{k-1}, Z_k \geq u_k, \tilde{Z}_k \geq c_k; \theta), \\ \xi_k(l_1, \ldots, l_k, u_k, c_k; \theta) &= \mathbb{P}(Z_1 \in \mathcal{C}_1, \ldots, Z_{k-1} \in \mathcal{C}_{k-1}, Z_k \leq l_k, \tilde{Z}_k \geq c_k; \theta).\end{aligned}$$

Then, under our given choice of decision constants, the type I error rate at $\theta = 0$ of our new test is

$$\begin{aligned}\mathbb{P}(\text{Reject } H_0; \theta = 0) &= \sum_{k=1}^{K-1} (\psi_k(l_1, \ldots, l_k, u_k, c_k; 0) + \xi_k(l_1, \ldots, l_k, u_k, c_k; 0)) \\ &\quad + \mathbb{P}(Z_1 \in \mathcal{C}_1, \ldots, Z_{K-1} \in \mathcal{C}_{K-1}, Z_K \geq u_K; \theta = 0) \\ &= \sum_{k=1}^{K} \mathbb{P}(Z_1 \in \mathcal{C}_1, \ldots, Z_{k-1} \in \mathcal{C}_{k-1}, Z_k \geq u_k; \theta = 0), \qquad (5.13)\end{aligned}$$

where the RHS of (5.13) is equal to α by definition of the critical values of the original standard GST. Referring to Table 4.10, we see that under $\theta = \delta$, the reversal probabilities of our optimal delayed response designs are such that

$$\begin{aligned}\mathbb{P}(Z_1 \in \mathcal{C}_1, \ldots, Z_{k-1} \in \mathcal{C}_{k-1}, Z_k \geq u_k, \tilde{Z}_k < c_k; \theta = \delta) \\ < \mathbb{P}(Z_1 \in \mathcal{C}_1, \ldots, Z_{k-1} \in \mathcal{C}_{k-1}, Z_k \leq l_k, \tilde{Z}_k \geq c_k; \theta = \delta). \qquad (5.14)\end{aligned}$$

This result says that if the true underlying value of θ is positive then the probability of a correct switch is greater than the probability of an incorrect one. If this holds for

our adapted GSTs for delayed responses, then

$$\psi_k(l_1,\ldots,l_k,u_k,c_k;\delta)+\xi_k(l_1,\ldots,l_k,u_k,c_k;\delta) > \mathbb{P}(Z_1\in\mathcal{C}_1,\ldots,Z_{k-1}\in\mathcal{C}_{k-1},Z_k\geq u_k;\theta=\delta),\quad \text{for } 1,\ldots,K-1,$$

and the power of the new test will be greater than $1-\beta$ under $\theta=\delta$.

In this section, we have shown how to formulate group sequential designs for delayed responses that can be planned ahead of time based on applying the stopping rule of a standard GST at each interim analysis. Referring to Section 4.4, where we list the expected sample sizes of standard GSTs when response is delayed, we see that for small values of r, the resultant tests are likely to be reasonably efficient. When $r=0.1$, applying the stopping rule of a standard GST optimal for F_2 when response is immediate, we lose less than 1% of n_{fix} compared with if we used an optimal delayed response GST. However, for larger values of r, using tests optimal specifically for delayed responses can make us extra savings of almost 10% when $r=0.4$.

5.7.2 Inference after a group sequential test has unexpectedly overrun

Suppose that a clinical trial is conducted according to an immediate response design. Subject responses to treatment can be observed almost instantaneously upon commencement of treatment and so overrunning is not anticipated at the design stage. However, when the trial is conducted, there is a delay in transferring and cleaning the data set ready for the interim analysis, during which recruitment continues. The data from these additional subjects must be incorporated into any decision analysis. The question posed is how one should "rescue" the analysis of such a test where overrunning was not envisaged ahead of time? In this section, we propose methodology for addressing this issue.

In Section 5.7.1, an immediate response GST was adapted in order to plan ahead for the fact that in the advent of early stopping at an interim analysis, the test would overrun. To recap, at decision analysis k, we find a constant c_k satisfying equation (5.7) such that we reject H_0 if $\tilde{Z}_k\geq c_k$, and accept H_0 otherwise. Under this construction, the resultant test's type I error probability under $\theta=0$, given in (5.13), does not depend on the value of $\tilde{I}_k-I_k$, for $k=1,\ldots,K-1$. Hence, the value of $\tilde{I}_k-I_k$, for $k=1,\ldots,K-1$, need not be known in advance; so long as the information levels $\{I_1,\ldots,I_K\}$ are attained, the type I error rate constraint will be satisfied under any observed sequence of overruns $\tilde{I}_1-I_1,\ldots,\tilde{I}_{K-1}-I_{K-1}$. In addition, the one-sided upper p-value for testing $H_0:\theta=0$ based on the ordering (5.8) does not depend on

the extent to which the test would have overrun if the test had stopped at an earlier interim analysis. The above properties mean that our method has great flexibility. The tests proposed in Section 5.7.1 can be implemented and p-values calculated in the case of unexpected overrunning where, supposing the test terminates at stage $k^\star$, the value of $\tilde{I}_k - I_k$, for $k = 1, \ldots, k^\star - 1, k^\star + 1, \ldots, K - 1$, will not be known. For example, suppose we close recruitment at interim analysis $k^\star$ only to find unexpectedly that there are subjects in the pipeline. Following the methods of Section 5.7.1, we can act as if we had planned for this overrun from the outset; we find the constant $c_{k^\star}$ satisfying the symmetry criterion (5.7) such that at the decision analysis we reject H_0 if $\tilde{Z}_{k^\star} \geq c_k$ and accept H_0 otherwise. P-values for testing H_0 are then calculated following the proposed ordering of Section 5.4.

When addressing the question of unexpected overrunning there is also a basic problem with deciding on a method of inference, e.g., ordering of the sample space, etc. after seeing the data; there is a danger that we may "shop around" for orderings which lead to more impressive p-values. To circumvent this problem, it would therefore be preferable to have a background policy on "what to do if there is overrunning" and then apply this if it is necessary. One can see that under the proposed strategy, things simply resolve back to the "immediate response" GST is there is no overrunning.

5.8 Appendix

Theorem 3. *$(\tilde{I}_T, \tilde{Z}_T)$ is a sufficient statistic for θ on termination of a GST for delayed responses.*

Proof: Consider a K-stage delayed response GST generating the sequence of standardised test statistics $\{Z_1, \tilde{Z}_1, \ldots, Z_{K-1}, \tilde{Z}_{K-1}, \tilde{Z}_K\}$ based on the information sequence $\{I_1, \tilde{I}_1, \ldots, I_{K-1}, \tilde{I}_{K-1}, \tilde{I}_K\}$. The GST stops at stage $T = \min\{k : Z_k \notin \mathcal{C}_k\}$, where $\mathcal{C}_k$ is the continuation region at interim analysis k and we define $\mathcal{C}_K = \emptyset$. Define the vector $\mathbf{Z}^{(k)} = (Z_1, \ldots, Z_k, \tilde{Z}_k)$ and, for each $k = 1, \ldots, K$, define

$$\mathcal{A}_k = \{\mathbf{z}^{(k)} : \ z_i \in \mathcal{C}_i,\ i = 1, \ldots, k-1,\ z_k \notin \mathcal{C}_k\},$$

the set of sample paths for $Z_1, Z_2, \ldots$ which first exit the continuation region at interim analysis k. The subsequence $(Z_1, \ldots, Z_k, \tilde{Z}_k)$ can be represented in terms of independent normal random variables. To see this, let $\Delta_i = I_i - I_{i-1}$, $i = 2, \ldots, k$, and $\tilde{\Delta}_k = \tilde{I}_k - I_k$. Suppose we have independent random variables $Y_1 \sim N(I_1\theta, I_1)$,

$Y_i \sim N(\Delta_i\theta, \Delta_i)$, $i = 2, \ldots, k$, and $\tilde{Y}_k \sim N(\tilde{\Delta}_k\theta, \tilde{\Delta}_k)$. Then, the sequence

$$\bigcup_{j=1}^{k} \left\{ \frac{1}{\sqrt{I_j}} \sum_{i=1}^{j} Y_i \right\} \cup \left\{ (Y_1 + \ldots + Y_k + \tilde{Y}_k)/\sqrt{\tilde{I}_k} \right\}$$

has the same joint distribution as $(Z_1, \ldots, Z_k, \tilde{Z}_k)$. In view of this representation, we can write the joint density of $(Z_1, \ldots, Z_k, \tilde{Z}_k)$, $p_{T,\tilde{Z}_T}(k, \mathbf{z}^{(k)}; \theta)$, at a point $\mathbf{z}^{(k)} \in \mathcal{A}_k$, so that $T = k$, as

$$p_{T,\tilde{Z}_T}(k, \mathbf{z}^{(k)}; \theta) = \left[\prod_{i=1}^{k} \frac{\sqrt{I_i}}{\sqrt{\Delta_i}} \phi\left(\frac{y_i - \Delta_i\theta}{\sqrt{\Delta_i}} \right) \right] \frac{\sqrt{\tilde{I}_k}}{\sqrt{\tilde{\Delta}_k}} \phi\left(\frac{\tilde{y}_k - \tilde{\Delta}_k\theta}{\sqrt{\tilde{\Delta}_k}} \right),$$

where $y_1 = z_1\sqrt{I_1}$, $y_i = z_i\sqrt{I_i} - z_{i-1}\sqrt{I_{i-1}}$, $i = 2, \ldots, k$, and $\tilde{y}_k = \tilde{z}_k\sqrt{\tilde{I}_k} - z_k\sqrt{I_k}$.

For a fixed k, $\tilde{Z}_k \sim N(\theta\sqrt{\tilde{I}_k}, 1)$ and so the likelihood of the data $L(\theta; \tilde{z_k})$ is proportional to $\exp\{\theta\sqrt{\tilde{I}_k}\tilde{z}_k - \theta^2\tilde{I}_k/2\}$. Hence, it comes as no surprise that for our problem we obtain

$$p_{T,\tilde{Z}_T}(k, \mathbf{z}^{(k)}; \theta) = g(k, \mathbf{z}^{(k)}, I_1, \ldots, I_k, \tilde{I}_k) \exp\{\theta\sqrt{\tilde{I}_k}\tilde{z}_k - \theta^2\tilde{I}_k/2\},$$

where

$$g(k, \mathbf{z}^{(k)}, I_1, \ldots, I_k, \tilde{I}_k) = \left[\prod_{i=1}^{k} \frac{\sqrt{I_i} \exp\{-y_i^2/(2\Delta_i)\}}{\sqrt{2\pi\Delta_i}} \right] \frac{\sqrt{\tilde{I}_k} \exp\{-\tilde{y}_k^2/(2\tilde{\Delta}_k)\}}{\sqrt{2\pi\tilde{\Delta}_k}}.$$

Note that y_i and $\tilde{y}_k$ are functions of $\mathbf{z}^{(k)}$ and $I_1, \ldots, I_k, \tilde{I}_k$ that do not involve θ. Since $T = k$, we can write

$$p_{T,\tilde{Z}_T}(k, \mathbf{z}^{(k)}; \theta) = g(k, \mathbf{z}^{(k)}, I_1, \ldots, I_k, \tilde{I}_k) \exp\{\theta\sqrt{\tilde{I}_T}\tilde{z}_T - \theta^2\tilde{I}_T/2\},$$

and, by the Neyman factorisation theorem, we conclude that the pair $(\tilde{I}_T, \tilde{Z}_T)$ is a sufficient statistic on termination for θ, as required. □

Chapter 6

Error spending tests for delayed responses

Suppose we wish to test the null hypothesis $H_0 : \theta = 0$ against the one-sided alternative $H_1 : \theta > 0$ with type I error probability α at $\theta = 0$ and power $1 - \beta$ at $\theta = \delta$. The delayed response GSTs of Chapters 2 and 3 were designed for K groups of subjects giving rise to information levels $\{I_1, \tilde{I}_1, \ldots, I_{K-1}, \tilde{I}_{K-1}, \tilde{I}_K\}$, where K is to be fixed at the design stage. In practice, varying accrual rates may mean that there is some deviation from this information sequence when the test is carried out which in turn can lead to changes in a test's error rates, assuming, that is, that the test can still be applied. In this chapter, we show that one can extend the methodology of Chapter 2 to deal with unpredictable sequences of information. Our approach is to derive error spending tests which attain nominal type I error rates exactly under any observed sequence of information levels. The development of these methods signals the culmination of the work presented in the early part of this thesis in a group sequential testing approach for delayed responses which can be easily implemented in practice.

6.1 Introduction

6.1.1 The "error spending" concept

For ease of presentation, we follow the historical development of error spending tests and first consider methodology for two-sided "immediate response" tests of $H_0 : \theta = 0$ when we have unpredictable sequences of information. The Wang & Tsiatis (1987) family of two-sided tests, of which the Pocock (1977) and O'Brien & Fleming (1979) tests are special cases, assume that K equally sized groups will be observed, where

both K and the group size are fixed at the design stage. Jennison & Turnbull (2001) extend the scope of these tests to unequally grouped data using a "significance level" approach. The test stops to reject H_0 at stage $k = 1, \ldots, K$, if the non-sequential test based on $|Z_k|$ rejects H_0 at two-sided significance level $2\{1-\Phi(c_k)\}$. Since only marginal significance levels are controlled at each stage, ignoring the possibility of early stopping, type I error rates are controlled only approximately unless observed information levels remain equally spaced. Jennison & Turnbull (2000, Section 3.3) investigate the attained operating characteristics of two-sided tests adapted using the significance level approach and find that large perturbations from the anticipated sequence of information levels can lead to changes in error rates which are undesirable.

Slud & Wei (1982) first introduced the concept of "error spending" for two-sided tests. The maximum number of analyses, K, and constants $\pi_1, \ldots, \pi_K$ are fixed at the design stage so that $\pi_1 + \ldots + \pi_K = \alpha$. At stage k, information levels $I_1, \ldots, I_k$ will have been observed. Given critical values $c_1, \ldots, c_{k-1}$, c_k is found as the solution to

$$\mathbb{P}(|Z_1| < c_1, \ldots, |Z_{k-1}| < c_{k-1}, |Z_k| \geq c_k; \theta = 0) = \pi_k. \tag{6.1}$$

We can think of π_k as the type I error probability to be spent at stage k, where under this construction the test will attain its type I error rate exactly under any observed sequence of information levels $I_1, \ldots, I_K$.

Conditional on the observed sequence of information levels $\{I_1, \ldots, I_K\}$, the sequence of test statistics $\{Z_1, \ldots, Z_K\}$ follows the usual canonical distribution if I_{k+1} is conditionally independent of $\{\hat{\theta}_1, \ldots, \hat{\theta}_k\}$ given $\{I_1, \ldots, I_k\}$ for each $k = 1, \ldots, K-1$. It follows that, in this case the probability on the LHS of (6.1) can be calculated as proposed in Section 2.2.2. We shall assume such conditional independence when we discuss error spending tests. This assumption remains valid if I_{k+1} is a function only of previous information levels. Hence, one may plan future recruitment in response to current accrual patterns and, given $\{I_1, \ldots, I_K\}$, $\{Z_1, \ldots, Z_K\}$ will still follow the usual canonical joint distribution. However, conditional independence will be violated if I_{k+1} is specified in response to $\hat{\theta}_k$; if error spending boundaries derived assuming independence are used to monitor the sequence of test statistics thus generated, the type I error rate of the test will be inflated. Other methods, known as adaptive designs, make a feature of allowing sample size to be re-estimated on the basis of updated estimates of θ but these need special construction in order to control the type I error rate.

The maximum information two-sided error spending tests of Lan & DeMets (1983)

address the limitations that K and $\pi_1, \ldots, \pi_K$ be fixed in advance in the method of Slud & Wei. Instead they stipulate that we fix in advance a target maximum information level, I_{max}, and a non-decreasing function f satisfying $f(0) = 0$ and $f(t) = \alpha$ for $t \geq 1$. The error spending function f controls how type I error probability is spent in response to observed information levels; $f(t)$ is the cumulative type I error probability spent when a fraction t of I_{max} has been accrued. For each $k = 1, 2, \ldots$, let $t_k = I_k / I_{max}$. Type I error probability

$$\begin{aligned} \pi_1 &= f(t_1) \\ \pi_k &= f(t_k) - f(t_{k-1}) \qquad\qquad k = 2, 3, \ldots, \end{aligned}$$

is spent at analyses $1, 2, \ldots$. Then, just as in Slud & Wei, given information levels $I_1, \ldots, I_k$, the critical value c_k is found as the solution to (6.1).

The test is designed under the assumption that unless the stopping rule directs us otherwise, sampling continues until the target information level has been reached. The test is then terminated at the first analysis for which $t_k \geq 1$. Let K be the maximum number of analyses that we permit in a particular realisation of the test. For a particular choice of f, the value of I_{max} can then be chosen so that under K equally spaced interim analyses, the test has power $1 - \beta$ at $\theta = \pm\delta$. One may also set a limit, ahead of time, of K analyses to be conducted upon implementation of the test, whatever the information sequence observed. Then, it is possible when the test is conducted that $I_K < I_{max}$, and the test is said to have underrun; the remainder of the type I error probability is spent at this final analysis and the power absorbs any perturbations. The final information level reached is highly influential on the power attained (Jennison & Turnbull (2000, Chapter 7)); the difference in I_K versus I_{max} will be commensurate with the perturbation in power.

For the remainder of this chapter, we focus our attention on one-sided testing problems. Suppose we wish to test $H_0 : \theta = 0$ against $H_1 : \theta > 0$ with type I error probability α and power $1 - \beta$ at $\theta = \delta$. Functions f and g are chosen to spend the type I and type II error probabilities, respectively. These functions must be non-decreasing, satisfying $f(0) = g(0) = 0$, $f(t) = \alpha$ and $g(t) = \beta$ for $t \geq 1$. At each stage $k = 1, 2, \ldots$, type I error probability under $\theta = 0$, $\pi_{1,k}$, and type II error probability under $\theta = \delta$, $\pi_{2,k}$, is spent, where

$$\begin{aligned} \pi_{1,1} &= f(t_1) & \pi_{2,1} &= g(t_1) \\ \pi_{1,k} &= f(t_k) - f(t_{k-1}) & \pi_{2,k} &= g(t_k) - g(t_{k-1}) \qquad k = 2, 3, \ldots \end{aligned}$$

As the test proceeds and information levels $I_1, \ldots, I_k$ are observed, we search to find

the critical values l_k and u_k satisfying

$$\mathbb{P}(l_1 < Z_1 < u_1, \ldots, l_{k-1} < Z_{k-1} < u_{k-1}, Z_k \geq u_k; \theta = 0) = \pi_{1,k},$$
$$\mathbb{P}(l_1 < Z_1 < u_1, \ldots, l_{k-1} < Z_{k-1} < u_{k-1}, Z_k \leq l_k; \theta = \delta) = \pi_{2,k}.$$

In designing the study, a search is conducted to find the value of I_{max} under which the test with K equally spaced analyses and a given sequence of $\pi_{1,k}$ and $\pi_{2,k}$ terminates properly at stage K with $l_K = u_K$. Values of I_{max} which are too small result in termination at stage K with $u_K > l_K$ whilst values which are too large mean our boundaries cross at stage $k < K$ with $l_k > u_k$. Hence, the target value of I_{max} can be found using a bisection search. In the next section, we shall consider how the error spending functions f and g should be chosen.

6.1.2 Choice of error spending function

A simple family of one-sided error spending functions, referred to as the ρ-family by Jennison & Turnbull (2000, Section 7.3) is given by

$$f(t) = \alpha \min\{t^\rho, 1\}, \qquad g(t) = \beta \min\{t^\rho, 1\}, \tag{6.2}$$

where functions are indexed by the power $\rho > 0$. Another family of functions, adapted for one-sided tests by Chang et al. (1998) from Hwang et al. (1990) who worked with two-sided tests, is indexed by the parameter $\gamma \in \mathbb{R}$ and is given by

$$\begin{aligned}
&f(t) = \alpha \min\{t, 1\}, && g(t) = \beta \min\{t, 1\} && \text{for } \gamma = 0,\\
&f(t) = \alpha \left(\frac{1 - e^{-\gamma \min\{1,t\}}}{1 - e^{-\gamma}} \right), && g(t) = \beta \left(\frac{1 - e^{-\gamma \min\{1,t\}}}{1 - e^{-\gamma}} \right) && \text{for } \gamma \neq 0.
\end{aligned}$$

We refer to this family as the γ-family of error spending functions. For both families of functions, the same proportion of the total type I and II error probabilities are spent when a fraction t of I_{max} has been accrued, so that $f(t)/\alpha = g(t)/\beta$, for all $t > 0$. For $\rho = 1$ and $\gamma = 0$, functions in the γ and ρ families coincide and error probabilities are spent at a constant rate as information accumulates. As ρ and γ vary, so do the operating characteristics of the tests found under the error spending functions they index. For example, for a given α and β, for small ρ, a more aggressive testing strategy is adopted with larger error probabilities being spent at lower information levels. The same is also true for tests in the γ-family when γ is large.

It is shown in Section 6.5 that for tests in the ρ-family, under the assumption of equally spaced analyses, the I_{max} required for a given K, α, β and ρ is proportional to $1/\delta^2$. This result has been stated by others, see for example Jennison & Turnbull (2000,

Section 7.3), but we prove it here for completeness. Hence, I_{max} can be written as $R_\rho(K, \alpha, \beta, \rho)I_{f,1}$, where $I_{f,1}$ is the corresponding one-sided fixed sample information level. The same also applies to tests in the γ-family and we write the information inflation factor as $R_\gamma(K, \alpha, \beta, \gamma)$. Under more aggressive testing strategies, we require higher values of I_{max} for the test to terminate properly at stage K with $l_K = u_K$ because more type I error probability is spent at lower information levels. Hence, we conjecture that the required information inflation factor $R_\rho(\alpha, \beta, K, \rho)$ is a decreasing function of ρ and $R_\gamma(\alpha, \beta, K, \gamma)$ is an increasing function of γ.

It is possible to extend the ideas outlined in this section to derive error spending tests for delayed responses which can cope with unpredictable sequences of information. In the next section, we shall explain in greater detail how exactly this can be done.

6.2 Error spending tests for delayed responses

6.2.1 Constructing error spending boundaries

Suppose we wish to test $H_0 : \theta = 0$ against the alternative $H_1 : \theta > 0$ using a delayed response GST. Accrual rates are subject to random variation, so we need the flexibility to be able to adapt to unpredictable sequences of information. The tests derived in this section assume that for each $k = 1, 2, \ldots$, $\tilde{I}_k$ is known at the time of interim analysis k. This seems reasonable since subjects are likely to be entered into a database upon entry into a study. Hence, it should be relatively easy to ascertain at any time the number of subjects recruited into a trial. In a similar spirit to (2.6)-(2.9), for $k = 1, 2, \ldots$, define

$$\begin{aligned}
&\psi_k(l_1, u_1, \ldots, l_{k-1}, u_{k-1}, l_k, u_k, c_k; \theta) \\
&\qquad = \mathbb{P}(l_1 < Z_1 < u_1, \ldots, l_{k-1} < Z_{k-1} < u_{k-1}, Z_k \geq u_k, \tilde{Z}_k \geq c_k; \theta) \\
&\qquad + \mathbb{P}(l_1 < Z_1 < u_1, \ldots, l_{k-1} < Z_{k-1} < u_{k-1}, Z_k \leq l_k, \tilde{Z}_k \geq c_k; \theta), \\
&\xi_k(l_1, u_1, \ldots, l_{k-1}, u_{k-1}, l_k, u_k, c_k; \theta) \\
&\qquad = \mathbb{P}(l_1 < Z_1 < u_1, \ldots, l_{k-1} < Z_{k-1} < u_{k-1}, Z_k \geq u_k, \tilde{Z}_k < c_k; \theta) \\
&\qquad + \mathbb{P}(l_1 < Z_1 < u_1, \ldots, l_{k-1} < Z_{k-1} < u_{k-1}, Z_k \leq l_k, \tilde{Z}_k < c_k; \theta).
\end{aligned}$$

Suppose we reach interim analysis k with $\tilde{I}_k < I_{max}$ and must decide whether or not to continue recruitment. At stage k, we wish to spend type I error probability under $\theta = 0$ $\pi_{1,k}$ and type II error probability under $\theta = \delta$ $\pi_{2,k}$. To do this, we must solve

the following pair of simultaneous equations for the boundary constants l_k, u_k and c_k:

$$\psi_k(l_1, u_1, \ldots, l_{k-1}, u_{k-1}, l_k, u_k, c_k; \theta = 0) = \pi_{1,k} \tag{6.3}$$

$$\xi_k(l_1, u_1, \ldots, l_{k-1}, u_{k-1}, l_k, u_k, c_k; \theta = \delta) = \pi_{2,k}. \tag{6.4}$$

We have two equations in three unknowns and therefore, without imposing any further constraints on our boundary constants, there is no unique solution for l_k, u_k and c_k. However, it is not entirely obvious what constraints we should impose on our test boundaries. To solve this problem, we shall look to the ideas of Section 5.7.1, developed in the context of adapting standard GSTs to deal with delayed responses. To see why this might help, suppose we have a fixed sequence of information levels $\{I_1, \ldots, \tilde{I}_K\}$. The boundary constants $\{l_1, u_1, \ldots, l_{K-1}, u_{K-1}, u_K\}$ define a standard GST of $H_0 : \theta \leq 0$ against $H_1 : \theta > 0$ based on this information sequence with type I error probability α at $\theta = 0$ and power $1 - \beta$ at $\theta = \delta$. However, when designing our test of H_0, we realise that there is a delay in response. We wish to adapt the standard GST to formulate a delayed response GST of H_0 based on the information sequence $\{I_1, \tilde{I}_1, \ldots, I_{K-1}, \tilde{I}_{K-1}, \tilde{I}_K\}$ with type I error rate α at $\theta = 0$. Our problem is how to choose constants $c_1, \ldots, c_{K-1}$ to partition the range of $\tilde{Z}_k$ at each $\tilde{I}_k$. It turns our that choosing c_k to satisfy the symmetry condition

$$\begin{aligned}&\mathbb{P}(l_1 < Z_1 < u_1, \ldots, l_{k-1} < Z_{k-1} < u_{k-1}, Z_k \geq u_k, \tilde{Z}_k < c_k; \theta = 0) \\ &\quad = \mathbb{P}(l_1 < Z_1 < u_1, \ldots, l_{k-1} < Z_{k-1} < u_{k-1}, Z_k \leq l_k, \tilde{Z}_k \geq c_k; \theta = 0),\end{aligned} \tag{6.5}$$

for $k = 1, \ldots, K-1$, ensures that the type I error rate at $\theta = 0$ of the delayed response GST based on boundary constants $\{l_1, u_1, c_1, \ldots, l_{K-1}, u_{K-1}, c_{K-1}, u_K\}$ is α, as required.

We adapt the ideas of Section 5.7.1 to solve our problem of how to choose the boundary constants l_k, u_k and c_k for our error spending test at stage k so that we spend the required error probabilities. We stipulate that c_k should be found as the solution to (6.5) so that the "switching" probabilities under $\theta = 0$ are equalised. Then, l_k and u_k can be found as the unique solutions to equations (6.3) and (6.4). If, for any given l_k and u_k, c_k is chosen to satisfy equation (6.5),

$$\psi_k(l_1, u_1, \ldots, l_k, u_k, c_k; \theta = 0) = \mathbb{P}(l_1 < Z_1 < u_1, \ldots, l_{k-1} < Z_{k-1} < u_{k-1}, Z_k \geq u_k; \theta = 0),$$

and we can write the type I error probability spent at stage k as a function of the unknown u_k only. Hence, finding u_k as the solution to

$$\mathbb{P}(l_1 < Z_1 < u_1, \ldots, l_{k-1} < Z_{k-1} < u_{k-1}, Z_k \geq u_k; \theta = 0) = \pi_{1,k},$$

which can be done using a bisection search over values of u_k, ensures we spend error probability $\pi_{1,k}$ at stage k. Fixing u_k, the decision constant c_k is a function of l_k only, and we write $c_k = h(l_k; u_k)$. We see that we have managed to "uncouple" our search for l_k and u_k; l_k is now found as the solution to

$$\xi_k(l_1, u_1, \ldots, l_k, u_k, h(l_k; u_k); \theta = \delta) = \pi_{2,k},$$

so that we spend type II error probability $\pi_{2,k}$ at stage k.

Recruitment is automatically terminated at the first interim analysis for which $\tilde{I}_k \geq I_{max}$. In this case, we do not analyse the data but instead wait to observe all recruited subjects before deciding whether or not to reject H_0. Spending all our remaining type I error probability, u_k is found as the solution to

$$\mathbb{P}(l_1 < Z_1 < u_1, \ldots, l_{k-1} < Z_{k-1} < u_{k-1}, \tilde{Z}_k \geq u_k; \theta = 0) = \alpha - \sum_{j=1}^{k-1} \pi_{1,j},$$

and we set $l_k = u_k$, rejecting H_0 if our test statistic at the final decision analysis $\tilde{Z}_k \geq u_k$ and accepting H_0 otherwise. This choice of l_k and u_k may mean that our test does not attain power exactly equal to $1 - \beta$ at $\theta = \delta$ but our priority is ensuring that the test terminates properly at the final stage with overall type I error rate α under $\theta = 0$.

6.2.2 Spending error probabilities

Note that Section 6.2.1 took the values $\pi_{1,k}$ and $\pi_{2,k}$ as a given. These should depend on the form of the error spending functions f and g. Barber & Jennison (2002) have shown that when response is immediate the γ and ρ families of error spending tests are highly efficient, in many cases performing close to optimal with respect to several criteria. In this section, we extend their efficiency results to our error spending tests for delayed responses and establish how robust their efficiency is to the delay parameter r. In particular, using the optimal delayed response GSTs derived in Chapter 3, we assess the efficiency of the γ and ρ families of error spending tests for delayed responses with respect to objective functions F_1 to F_4 as defined in Section 3.1.

$$F_1 = \mathbb{E}(N; \theta = \delta/2), \qquad F_2 = 0.5\{\mathbb{E}(N; \theta = 0) + \mathbb{E}(N; \theta = \delta)\},$$

$$F_3 = 0.5\{\mathbb{E}(N; \theta = -\delta/2) + \mathbb{E}(N; \theta = 3\delta/2)\}, \qquad F_4 = \int \mathbb{E}(N; \theta) \frac{2}{\delta} \phi\left(\frac{\theta - \delta/2}{\delta/2}\right) \mathrm{d}\theta.$$

Suppose we wish to design and implement an error spending test of $H_0 : \theta \leq 0$. For a given α, β, δ, K, Δ_t, R, c and σ^2 we fix the target information level $I_{max} = RI_{fix}$ and $t_{max} = 4\sigma^2 I_{max}/c$. This in turn fixes $r = \Delta_t/t_{max}$. Our test is going to have the form

$$
\begin{array}{ll}
\text{At interim analysis } k = 1, \ldots, K-2, & \\
\text{if } l_k < Z_k < u_k & \text{continue to interim analysis } k+1, \\
\text{otherwise} & \text{continue to decision analysis } k. \\
\text{At decision analysis } k = 1, \ldots, K-1, & \\
\text{if } \tilde{Z}_k \geq c_k & \text{reject } H_0, \\
\text{otherwise} & \text{accept } H_0. \\
 & \\
\text{At interim analysis } k = K-1, & \\
\text{if } l_k < Z_k < u_k & \text{continue to decision analysis } K, \\
\text{otherwise} & \text{continue to decision analysis } K-1. \\
\text{At decision analysis } K, & \\
\text{if } \tilde{Z}_k \geq u_K & \text{reject } H_0, \\
\text{if } \tilde{Z}_k < l_K & \text{accept } H_0.
\end{array}
\tag{6.6}
$$

The test is designed and implemented under the information sequence

$$
I_k = \frac{k}{K}(1-r)I_{max}, \qquad \tilde{I}_k = I_k + rI_{max}, \qquad k = 1, \ldots, K-1, \tag{6.7}
$$

where in the absence of early stopping, the test terminates at decision analysis K at information level $\tilde{I}_K = I_{max}$.

The error probabilities to be spent at each stage of our test should depend on the "information so far", but we have not yet specified just what this should mean. One possible strategy is to spend them as a function of the information levels $\tilde{I}_1, \ldots, \tilde{I}_K$, so that

$$
\begin{array}{ll}
\pi_{1,1} = f(\tilde{I}_1/I_{max}) & \pi_{2,1} = g(\tilde{I}_1/I_{max}) \\
\pi_{1,k} = f(\tilde{I}_k/I_{max}) - f(\tilde{I}_{k-1}/I_{max}) & \pi_{2,k} = g(\tilde{I}_k/I_{max}) - g(\tilde{I}_{k-1}/I_{max}) \qquad k = 2, \ldots, \mathrm{K}.
\end{array}
$$

We shall refer to this later as strategy 1. Functions f and g are chosen from the ρ-family of error spending functions defined in (6.2). We search for the value of $\rho(\alpha, \beta, K, R, r)$ to go with our problem, i.e., the value determining the sequence of probabilities $\pi_{1,k}$ and $\pi_{2,k}$ such that the corresponding test based on the information sequence $\{I_k, \tilde{I}_k\}$ terminates with $l_K = u_K$. Once this test has been found, it is evaluated for $F_1, \ldots, F_4$ and the values attained expressed as a percentage of the fixed sample size n_{fix}; these results are invariant to changes in Δ_t, c, σ^2 and δ so long as r stays fixed. This process

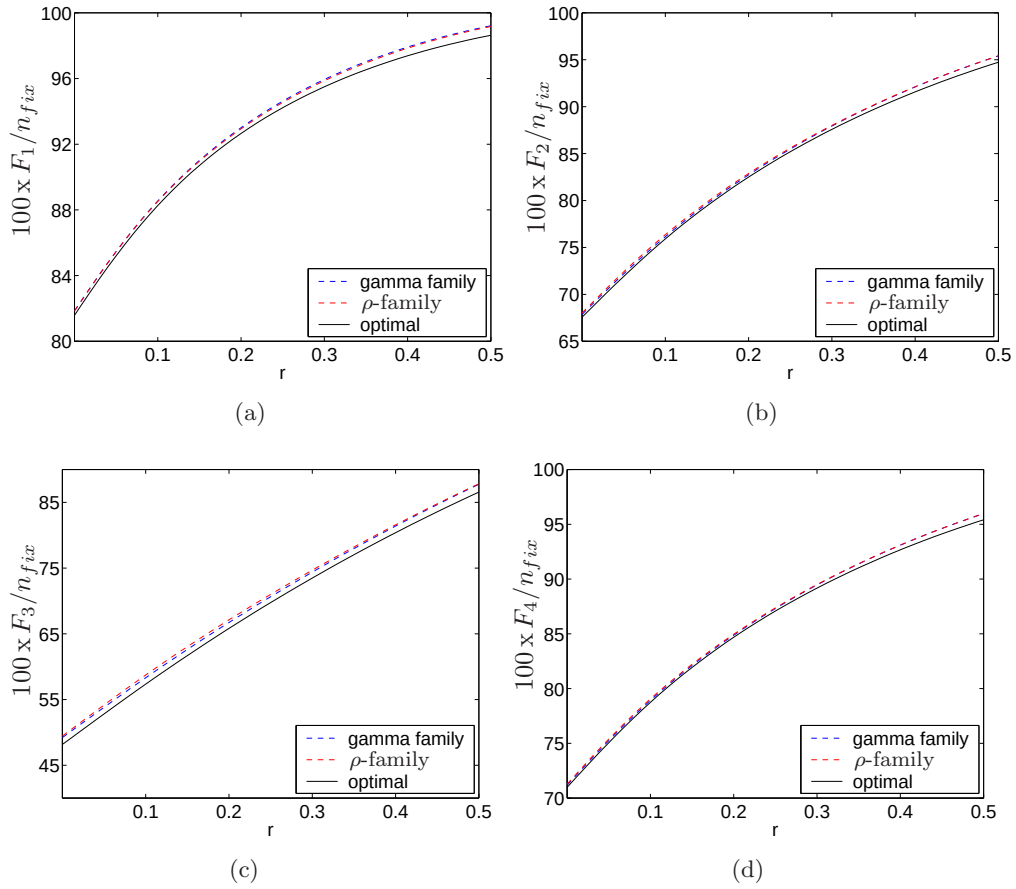


Figure 6-1: Objective functions attained by optimal delayed response GSTs and γ and ρ families of error spending tests for various values of the delay parameter r. Tests are designed and implemented under $\alpha = 0.05$, $\beta = 0.1$, $R = 1.15$ and $K = 3$ and error probabilities are spent as a function of $\tilde{I}_k$.

is repeated for the case when error spending functions are chosen from the γ-family. Figures 6-1(a) - 6-1(d) compare the values of the objective functions achieved by our error spending tests with the performances of optimal GSTs for a range of values of r when $K = 3$, $\alpha = 0.05$, $\beta = 0.1$ and $R = 1.15$. These efficiency results are certainly impressive; the error spending tests from both families are within 1% or 2% of n_{fix} of the optimal tests over the range of values of r considered.

Figures 6-1(a) - 6-1(d) show that there is little to choose between tests designed under the γ and ρ families of error spending functions. Indeed, the curves for these two families of tests lie almost on top of each other. In the case when $K = 2$, the curves for the γ and ρ-family of tests will lie exactly on top of each other. This is because for a given sequence of information levels $\{I_1, \tilde{I}_1, \tilde{I}_2\}$, two-stage tests constructed under γ and ρ error spending functions will be defined by the same boundary constants. To explain this, recall that these functions spend error probabilities symmetrically so that

at each stage k, $\pi_{1,k}/\alpha = \pi_{2,k}/\beta$. Once we have chosen the proportion of α and β to be spent at the first stage, we fix the remaining error probability to be spent at the final decision analysis. Hence, without loss of generality, the proportion $\pi_{1,1}/\alpha$ will control the boundary constants defining the test, namely $\{u_1, l_1, c_1, u_2, l_2\}$. There will be a unique pair of stage 1 error probabilities satisfying $\pi_{1,1}/\alpha = \pi_{2,1}/\beta$ such that the test terminates with $l_2 = u_2$ and hence the γ and ρ tests must be defined by the same boundary constants.

An alternative way of spending error probabilities, which we shall refer to as strategy 2, is in response to observed information levels $I_1, \ldots, I_{K-1}, \tilde{I}_K$, so that

$$\begin{aligned} \pi_{1,1} &= f(I_1/I_{max}) & \pi_{2,1} &= g(I_1/I_{max}), \\ \pi_{1,k} &= f(I_k/I_{max}) - f(I_{k-1}/I_{max}) & \pi_{2,k} &= g(I_k/I_{max}) - g(I_{k-1}/I_{max}) \quad k = 2, \ldots, K-1, \end{aligned}$$

and

$$\pi_{1,K} = f(\tilde{I}_K/I_{max}) - f(I_{K-1}/I_{max}) \qquad \pi_{2,K} = g(I_K/I_{max}) - g(I_{K-1}/I_{max}).$$

In practice, one may prefer to adopt this approach; one may be wary of spending error probabilities based on recruited information levels since only some of this information will be available at the crucial point when deciding whether to stop recruitment. Figures 6-2(a) - 6-2(d) compare the performances of ρ-family tests when error probabilities are spent according to strategies 1 and 2 for the case when $K = 3$, $R = 1.05$, $\alpha = 0.05$ and $\beta = 0.1$. We see that for objective functions F_1, F_2 and F_4, both strategies perform similarly; error spending tests are within around 1% of n_{fix} of the optimal test over the range of r considered. Figures 6-3(a) - 6-3(d) show that there is a greater disparity between the two methods of spending error probabilities when $K = 5$ and $R = 1.05$, but both methods are still highly efficient. For example, when $r = 0.25$, one can save 1% more on n_{fix} for F_4 by switching from spending error probabilities according to strategy 1 to strategy 2.

In the next section, we shall illustrate the error spending methodology for delayed responses developed so far by means of an example.

6.2.3 An illustrative example

Facey (1992), and later Mehta & Tsiatis (2001), cite a placebo controlled trial testing the efficacy of a new treatment intended to treat hypercholesterolemia. The primary endpoint is reduction in total serum cholesterol over 4 weeks. Observations $X_{A,i} \sim N(\mu_A, \sigma^2)$ and $X_{B,i} \sim N(\mu_B, \sigma^2)$, $i = 1, 2, \ldots$, are made on the new treatment

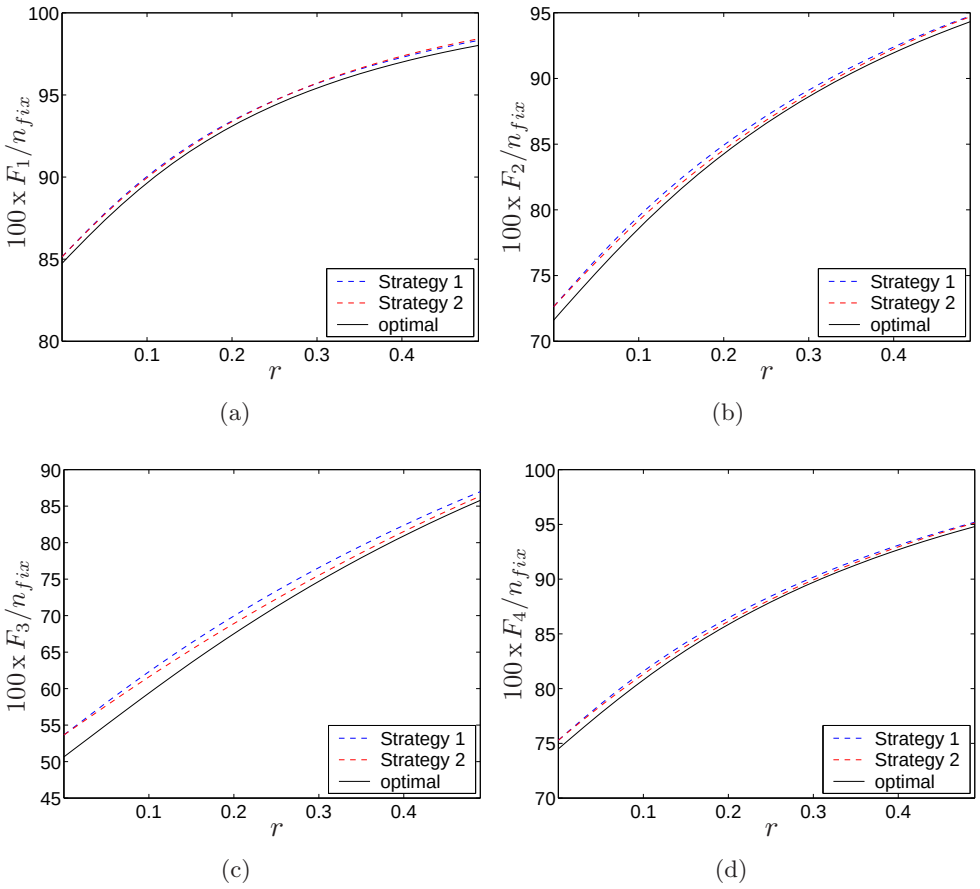


Figure 6-2: Objective functions attained by optimal delayed response GSTs and ρ families of error spending tests when error probabilities are spent according to strategies 1 and 2, i.e., as a function of $\{\tilde{I}_1, \ldots, \tilde{I}_K\}$ and $\{I_1, \ldots, I_{K-1}, \tilde{I}_K\}$, respectively. Tests are designed and implemented under $\alpha = 0.05$, $\beta = 0.1$, $R = 1.05$ and $K = 3$.

and control respectively. All observations are independent and we deviate slightly from the details given in Facey and assume it is known $\sigma^2 = 1$. Define $\theta = \mu_A - \mu_B$. We test $H_0 : \theta \leq 0$ against $H_1 : \theta > 0$ with type I error rate 0.05 at $\theta = 0$ and power 0.9 to detect a reduction in serum cholesterol of 1 mmol/litre by the new treatment. The corresponding fixed sample test requires information

$$I_{fix} = \{\Phi^{-1}(0.95) + \Phi^{-1}(0.9)\}^2 = 8.564.$$

Data are to be monitored group sequentially using error spending boundaries. We decide to fix the form of the error spending functions first and then find our target information level I_{max}. Following the discussion in Section 6.2.2, we select a pair of error spending functions from the ρ-family which will spend error probabilities as a function of $\tilde{I}_k$, the total information recruited at the interim analysis, rather than I_k.

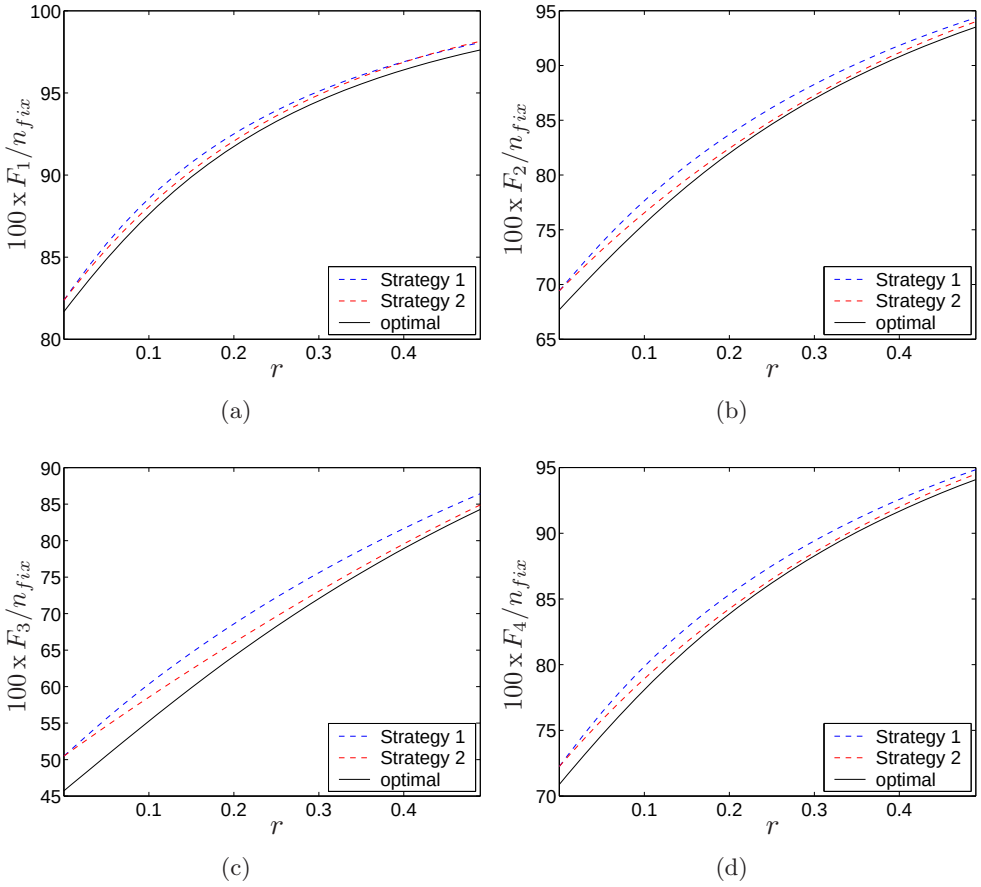


Figure 6-3: Objective functions attained by optimal delayed response GSTs and ρ families of error spending tests. The error probabilities are spent according to strategies 1 and 2, i.e., as a function of $\{\tilde{I}_1, \ldots, \tilde{I}_K\}$ and $\{I_1, \ldots, I_{K-1}, \tilde{I}_K\}$, respectively. Tests are designed and implemented under $\alpha = 0.05$, $\beta = 0.1$, $R = 1.05$ and $K = 5$.

Setting $\rho = 2$,

$$f(t) = 0.05 \min\{t^2, 1\} \quad \text{and} \quad g(t) = 0.1 \min\{t^2, 1\}, \tag{6.8}$$

so that error probabilities are spent slowly in the early stages of the test.

We set I_{max} so that a particular realisation of a $K = 3$ stage test of H_0 of the form (6.6) will terminate properly at the final stage with $l_K = u_K$. We design for the scenario in which recruitment occurs at a constant rate of 4 subjects per week; at each point in time patient entry is equally divided between treatments A and B. In the pipeline at each interim analysis there will be $c\Delta_t = 16$ subjects, 8 on each treatment, and $\tilde{I}_k - I_k = 16/4\sigma^2 = 4$ units of information. For a given $I_{max} = RI_{fix}$, interim and decision analyses are scheduled at information levels

$$I_k = \frac{k}{3}(I_{max} - 4) \quad \text{and} \quad \tilde{I}_k = I_k + 4, \quad k = 1, 2,$$

with the test terminating at information level $\tilde{I}_3 = I_{max}$. A numerical search finds that we should set our target information level as $I_{max} = 10.247$.

At the first scheduled interim analysis, we observe $I_1 = 1$ and standardised test statistic $Z_1 = 2$. We also observe $\tilde{I}_1 = 5$. Substituting the information ratio $\tilde{I}_1/I_{max} = 0.488$ into error spending functions (6.8), we calculate our boundary constants at this first stage to spend type I error probability $\pi_{1,1} = 0.00935$ and type II error probability $\pi_{2,1} = 0.0119$. Solving for u_1 the equation

$$\mathbb{P}(Z_1 \geq u_1; \theta = 0) = 0.00935,$$

we find $u_1 = 2.260$. Given this upper boundary constant, we search over values of l_1 to find the solution to

$$\mathbb{P}(Z_1 \geq 2.260, \tilde{Z}_1 < c_1; \theta = 1) + \mathbb{P}(Z_1 \leq l_1, \tilde{Z}_1 < c_1; \theta = 1) = 0.0119, \qquad (6.9)$$

where, for each value of l_1, c_1 is chosen to satisfy

$$\mathbb{P}(Z_1 \geq 2.260, \tilde{Z}_1 < c_1; \theta = 0) = \mathbb{P}(Z_1 \leq l_1, \tilde{Z}_1 \geq c_1; \theta = 0).$$

Solving (6.9) numerically, we find $l_1 = -0.688$ and $c_1 = 1.219$. We see the sample path lies within the continuation region at the first interim analysis; recruitment is continued and we progress to the next stage.

At the second interim analysis, we observe $I_2 = 5.5$ and $Z_2 = 1.6$ and also $\tilde{I}_2 = 9.5$. Following the methodology outlined in Section 6.2.1, we find boundary constants $l_2 = 1.470$, $u_2 = 1.812$ and $c_2 = 1.705$ which spend error probabilities $\pi_{1,2} = 0.0311$ and $\pi_{2,2} = 0.0621$ at this second stage. Again the sample path falls within the continuation region and sampling continues. At the third interim analysis, we observe $\tilde{I}_3 = 10.5 > I_{max}$ so we close recruitment without analysing the data. Once all pipeline subjects have been observed, we conduct a final analysis at information level $\tilde{I}_3$ and observe $\tilde{Z}_3 = 2.1$. All remaining error probabilities are spent so that $\pi_{1,3} = 0.00703$ and $\pi_{2,3} = 0.0141$. Solving the equation

$$\mathbb{P}(-0.688 < Z_1 < 2.260, 1.470 < Z_2 < 1.812, \tilde{Z}_3 \geq u_3; \theta = 0) = 0.00703,$$

for u_3, we find the solution is given by $u_3 = 1.712$ and set $l_3 = u_3$. We see that $\tilde{Z}_3 > u_3$ and so we terminate the test with rejection of H_0 and declare the new treatment as effective. The test has attained its overall type I error rate exactly, but the attained power is 0.91 at $\theta = 1$ due to deviations from the planned sequence of information levels. The small difference in $\tilde{I}_3$ compared to I_{max} is commensurate with the perturbation in

the power of the test.

6.3 Analysis on termination of an error spending test for delayed responses

In Section 5.4, we proposed a stage-wise ordering on the sample space defined by a delayed response GST. Conveniently, the p-value for testing H_0 based on this ordering did not depend on future information levels beyond the observed stopping stage. Hence, it is possible to calculate these p-values for testing H_0 on termination of an error spending test. We calculate p-values by conditioning on the observed information levels. This is reasonable since if information levels are not influenced by the values of preceding test statistics then I_k and $\tilde{I}_k$, $k = 1, 2, \ldots$, are ancillary statistics for θ. For example, consider the illustrative example of Section 6.2.3, where we observed $(T, \tilde{Z}_T) = (3, 2.1)$. Given information levels $I_1 = 1$, $\tilde{I}_1 = 5$, $I_2 = 5.5$, $\tilde{I}_2 = 9.5$ and $\tilde{I}_3 = 10.5$, the p-value for testing $H_0 : \theta = 0$ against $\theta > 0$ is

$$\begin{aligned} p^+ &= \sum_{k=1}^{2} \psi_k(l_1, u_1, \ldots, l_{k-1}, u_{k-1}, l_k, u_k, c_k; \theta = 0) \\ &\quad + \mathbb{P}(l_1 < Z_1 < u_1, l_2 < Z_2 < u_2, \tilde{Z}_3 \geq u_3; \theta = 0) \\ &= 0.0459. \end{aligned}$$

As expected, our p-value is below α which is consistent with the decision of the GST to reject H_0.

6.4 Error spending tests when the number of pipeline responses is unpredictable

The methods of Section 6.2 were based on the assumption that at each interim analysis $k = 1, 2, \ldots$, $\tilde{I}_k$ will be known. However, subject dropouts may mean that should recruitment be closed, less information will be available at the decision analysis than anticipated at the interim analysis. In this section, we extend the methodology of this chapter to derive error spending tests which give us the flexibility to deal with unpredictable amounts of overrun. We continue to follow the basic test structure of (6.6), but change the way we calculate error spending boundaries at an interim analysis.

At each interim analysis $k = 1, 2, \ldots$, we must calculate the type I and type II error probabilities to be spent at stage k. Since the value of $\tilde{I}_k$ is not known, we choose to

spend error probabilities as a function of $t_k = I_k/I_{max}$, so that

$$\pi_{1,1} = f(t_1) \qquad \pi_{2,1} = g(t_1)$$
$$\pi_{1,k} = f(t_k) - f(t_{k-1}) \qquad \pi_{2,k} = g(t_k) - g(t_{k-1}) \qquad \text{for } k = 2, \ldots.$$

Boundary constants l_k and u_k are then found as solutions to the pair of equations

$$\mathbb{P}(l_1 < Z_1 < u_1, \ldots, l_{k-1} < Z_{k-1} < u_{k-1}, Z_k \geq u_k; \theta = 0) = \pi_{1,k}, \tag{6.10}$$
$$\mathbb{P}(l_1 < Z_1 < u_1, \ldots, l_{k-1} < Z_{k-1} < u_{k-1}, Z_k \leq l_k; \theta = \delta) = \pi_{2,k}, \tag{6.11}$$

and so our critical values defining the continuation region at the interim analysis do not depend on future information levels beyond interim analysis k. Equations (6.10) and (6.11) can be solved using any standard software for computing the boundaries of an "immediate response" error spending test, such as EaSt (Cytel Software Corporation, 1992).

Should we choose to terminate recruitment, we must wait for all recruited subjects, minus any dropouts, to be observed before conducting a decision analysis. Following the methodology presented in Section 6.2.1, we find c_k so that it satisfies the symmetry criterion of (6.5). Under this choice of c_k, the type I error probability spent at stage k will be exactly equal to $\pi_{1,k}$, as intended. However, finding l_k as the solution to (6.11) means

$$\begin{aligned}\pi_{2,k} = {} & \mathbb{P}(l_1 < Z_1 < u_1, \ldots, l_{k-1} < Z_{k-1} < u_{k-1}, Z_k \leq l_k, \tilde{Z}_k < c_k; \theta = \delta) \\ & + \mathbb{P}(l_1 < Z_1 < u_1, \ldots, l_{k-1} < Z_{k-1} < u_{k-1}, Z_k \leq l_k, \tilde{Z}_k \geq c_k; \theta = \delta).\end{aligned}$$

Based on our results for optimal delayed response GSTs (see Table 4.10), we anticipate that, in general, the reversal probabilities for our adapted GSTs will be such that

$$\begin{aligned}&\mathbb{P}(l_1 < Z_1 < u_1, \ldots, l_{k-1} < Z_{k-1} < u_{k-1}, Z_k \leq l_k, \tilde{Z}_k \geq c_k; \theta = \delta) \\ &\quad > \mathbb{P}(l_1 < Z_1 < u_1, \ldots, l_{k-1} < Z_{k-1} < u_{k-1}, Z_k \geq u_k, \tilde{Z}_k < c_k; \theta = \delta),\end{aligned} \tag{6.12}$$

and this inequality has been found to hold empirically in the examples we have considered. It follows that

$$\pi_{2,k} > \xi_k(l_1, u_1, \ldots, l_{k-1}, u_{k-1}, l_k, u_k, c_k; \theta = \delta),$$

and we see that we will not manage to spend the necessary type II error probability at stage k in this case. Let $\tilde{n}_k$ be the number of recruited subjects at interim analysis k, excluding those who have previously dropped out. Recruitment is closed at the first analysis for which $\tilde{n}_k \geq n_{max}$, where n_{max} is the target sample size required to attain

		K			
	r	2	3	5	10
$R = 1.01$					
	0.01	0.900	0.900	0.900	0.900
	0.1	0.900	0.900	0.900	0.900
	0.2	0.900	0.900	0.900	0.900
$R = 1.15$					
	0.01	0.900	0.900	0.900	0.900
	0.1	0.901	0.901	0.901	0.902
	0.2	0.906	0.905	0.905	0.905
$R = 1.3$					
	0.01	0.900	0.900	0.900	0.900
	0.1	0.905	0.903	0.902	0.902
	0.2	0.914	0.911	0.910	0.910

Table 6.1: Power at $\theta = 1$ attained by ρ-family error spending tests of $H_0 : \theta = 0$ against $H_1 : \theta > 0$ designed to cope with unpredictable numbers of pipeline responses. Tests are designed and implemented under the sequence of information levels (6.7). Type I error probabilities at $\theta = 0$ and type II error probabilities at $\theta = 1$ are spent according to functions $f(t) = 0.05 \min\{t^\rho, 1\}$ and $g(t) = 0.1 \min\{t^\rho, 1\}$, where ρ is found as outlined on p.94.

information level I_{max}. We will subsequently fail to reach the target information level at the decision analysis if pipeline subjects drop out. As usual, at this final decision analysis we choose u_k so we spend all remaining type I error probability and then set $l_k = u_k$ to ensure the test terminates properly.

From Table 6.1, we see that for small values of the inflation factor R, the power attained by error spending tests designed to cope with unpredictable numbers of pipeline responses is close to $1-\beta$ at $\theta = \delta$. For larger values of r, the power achieved at this alternative increases. One possible explanation for this can be found by considering (6.12); the difference between attained power and $1 - \beta$ will increase with r if the LHS probability increases at a faster rate than the RHS probability. Looking at Figures 6-4(a) - 6-4(d), it is evident that this extra power comes at a price. For small $r < 0.3$, error spending tests continue to be efficient, but as r becomes larger their performance diverges from that of the optimal GSTs.

6.5 Appendix

Theorem 4. *The value of I_{max} required for a properly terminating error spending test of $H_0 : \theta = 0$ to attain power $1-\beta$ at $\theta = \delta$ and type I error probability α is proportional*

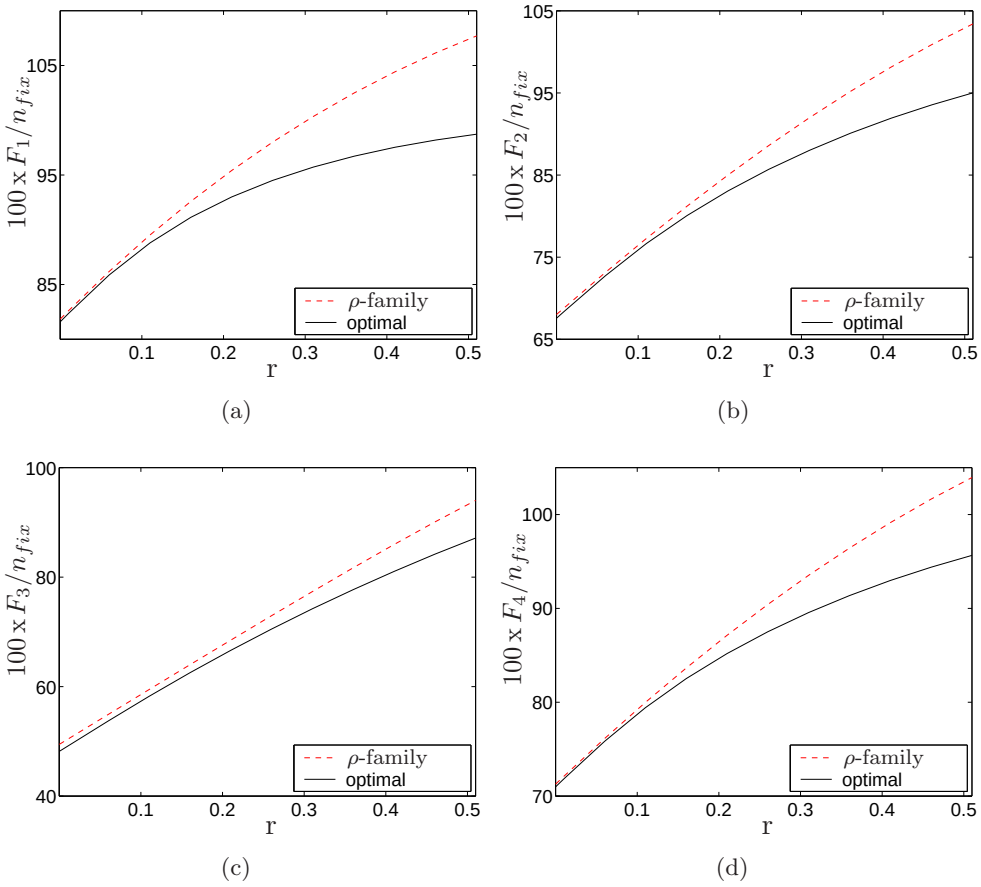


Figure 6-4: Objective functions attained by optimal delayed response GSTs and ρ-family error spending tests designed to cope with unpredictable numbers of pipeline responses. Tests are designed under $\alpha = 0.05$, $\beta = 0.1$, $R = 1.05$, and $K = 5$.

to $1/\delta^2$.

Proof: Suppose we are testing the null hypothesis $H_0 : \theta = 0$ against the two-sided alternative $H_1 : \theta \neq 0$. Setting

$$I_{max} = R\frac{\{\Phi^{-1}(1-\alpha/2) + \Phi^{-1}(1-\beta)\}^2}{\delta^2},$$

under the assumption of equally spaced analyses, $I_k/I_{k+1} = k/(k+1)$, for each $k = 1, \ldots, K-1$. Hence, under H_0, the joint distribution of the standardised statistics $Z_1, \ldots, Z_K$ does not depend on δ. Therefore, for a given α, β and K, the critical values $c_1, \ldots, c_K$ are a constant function of δ. The power of a test is the sum of K integrals where the integrands are functions of $f_1(z_1; \theta)$ and $f_k(z_k|z_{k-1}; \theta)$, where these are the marginal densities of Z_1 and $Z_k|Z_{k-1} = z_{k-1}$, for $k = 2, \ldots, K$ respectively. These

densities take the form shown below:

$$f_1(z_1;\theta) = \phi(z_1 - \theta\sqrt{(I_{max}/K)}) \tag{6.13}$$

$$f_k(z_k|z_{k-1};\theta) = \sqrt{k}\phi(z_k\sqrt{k} - z_{k-1}\sqrt{k-1} - \theta\sqrt{I_{max}/K}) \qquad \text{k=2,}\ldots,K \tag{6.14}$$

Note that θ only appears in these densities multiplied by the constant factor $\sqrt{I_{max}/K}$. We choose I_{max} so the power constraint at $\theta = \pm\delta$ is satisfied. Fixing α, β and K, and hence $\{l_1, u_1, \ldots, l_K, u_K\}$, there will be a unique value of $\delta\sqrt{I_{max}/K}$ such that this holds. Varying δ, it follows that we must have $I_{max} \propto 1/\delta^2$.

Turning our attention to one-sided tests, we consider testing $H_0 : \theta = 0$ against $H_1 : \theta > 0$. We search for the value of I_{max} such that the K-stage test terminates properly with $l_K = u_K$ having spent type I error probability α and type II error rate β at $\theta = \delta$. At each stage, error probabilities are spent as a function of the ratios I_k/I_{max}. Under equally spaced analyses, these will not depend on δ and so the shape of the boundaries is fixed as δ varies. A test's error probabilities are integrals of densities in the form (6.13) and (6.14). As δ varies, there will be a unique value of $\delta\sqrt{I_{max}/K}$ such that the K-stage test terminates properly at stage K with $l_K = u_K$; too high, we terminate at stage $k < K$ with $l_k > u_k$ and too low we terminate with $l_K < u_K$. It follows that I_{max} is proportional to $1/\delta^2$. □

Chapter 7

Short-term endpoints

The resounding message of Chapter 3 was that when designing GSTs for delayed responses, serious consideration should be given to how one can minimise the number of subjects in the pipeline at an interim analysis. For example, suppose we conduct a complex multi-centre study. The gains to be made by speeding up the collection and cleaning of data ready for an interim analysis will often merit the extra investment this requires. One must also re-think recruitment strategies; be aware that recruiting subjects as quickly as possible may not be best in terms of minimising a test's expected sample size. These conclusions echo comments made by Grieve & Krams (2005) when relecting upon the ASTIN study, an adaptive Phase II study based on a primary endpoint of response 90 days after treatment. Dose allocation ratios were adapted as information on the dose-response curve accumulated; if accrual was too quick, recruitment would be completed before there was time to learn much about the dose-response curve. For small values of the delay parameter r, the message from our results was promising: there are still good savings to be made on the fixed sample test using group sequential testing. However, as r increases to 0.4 and beyond, there are few remaining benefits of interim monitoring for early stopping.

The above results are derived assuming data will only be available on a primary long-term endpoint. However, observations are likely to be made across a wide spectrum of endpoints. Within this framework, measurements on a secondary short-term endpoint thought to be correlated with a subject's eventual long-term response could be made with little additional effort and incorporated into the stopping rule of a GST. Methodology for doing this was presented in Chapter 2. In this chapter, we extend this methodology to the case where the covariance matrix of the joint model for the primary and secondary endpoints isFor consistency with the notation of this chapter, we now refer to this as the λ family and index functions by the parameter $\lambda > 0$. unknown, and explore the gains to be made if a suitable short-term endpoint

can be found.

7.1 Methodology

Suppose we wish to compare a new experimental treatment against control. The new treatment will be deemed to be effective if it is shown to do better than control in some long-term indicator of treatment effect which can be measured after time Δ_t has elapsed. Observations on a correlated short-term endpoint will also be made after time $\Delta_{1,t} < \Delta_t$ has elapsed, but this endpoint may not be of direct clinical interest itself. Define $\kappa = \Delta_{1,t}/\Delta_t$ to be a measure of how quickly the secondary endpoint is available, where $\kappa \in [0, 1)$. Let $X_{A,i}$ and $X_{B,i}$, $i = 1, 2, \ldots$, denote the short-term responses of subjects allocated to two treatments, A and B. Let $Y_{A,i}$ and $Y_{B,i}$, $i = 1, 2, \ldots$, denote observations on the long-term endpoint for these subjects. Suppose responses on a subject can be modelled as

$$\begin{pmatrix} X_{T,i} \\ Y_{T,i} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{T,1} \\ \mu_{T,2} \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \tau\sigma_1\sigma_2 \\ \tau\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right) \quad \text{for } i = 1, 2, \ldots; T \in \{A, B\},$$

where for now we suppose σ_1^2, σ_2^2 and τ are known. Responses on different subjects are assumed to be independent. Let $\boldsymbol{\beta} = (\mu_{A,1}, \mu_{B,1}, \mu_{A,2}, \mu_{B,2})^T$ denote our vector of parameters, where inferences are to be made on $\theta = \mu_{A,2} - \mu_{B,2}$. We wish to test the null hypothesis $H_0 : \theta \leq 0$ against the one-sided alternative $H_1 : \theta > 0$ using a K-stage delayed response GST with type I error probability α at $\theta = 0$ and power $1 - \beta$ at $\theta = \delta$. We make the assumptions that throughout the trial, recruitment occurs at a constant rate such that at each point in time, equal numbers of subjects will have been randomised to each treatment.

At each interim analysis, of those pipeline subjects whose long-term response has yet to be observed, a certain fraction will be partially observed, with their short-term response available. For values of κ close to 0, this fraction will be close to 1, and the number of completely unobserved subjects will be small. A subject is said to be fully observed when both their short and long-term responses have been measured. For each $k = 1, \ldots, K-1$, let $n_{2,k}$ and $n_{1,k}$ denote the number of fully and partially observed subjects at interim analysis k when a total of $\tilde{n}_k$ have been recruited. Under the simplifying assumptions made about the pattern of recruitment into the trial, of those subjects fully and partially observed, half will have been randomised to each treatment. Figure 7-1 shows how the number of partially and fully observed subjects will vary during the course of the trial. Should recruitment be terminated at an interim analysis, we must wait until all recruited subjects are fully observed before deciding whether to reject or accept H_0.

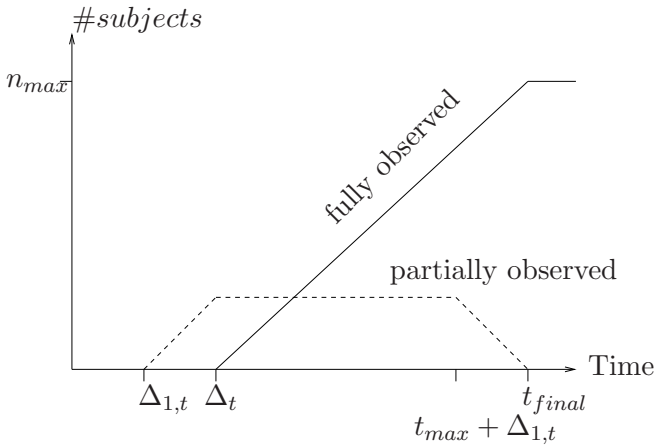


Figure 7-1: Illustration of how the numbers of fully and partially observed subjects vary during the course of a delayed response GST where a secondary endpoint is measured. During the interval $(\Delta_t, t_{max} + \Delta_{1,t})$, the total number of partially observed subjects remains constant at $n_{max} r(1-\kappa)$.

In order for θ to be estimable at an interim analysis, we require fully observed subjects on each treatment. Hence, the $K-1$ interim analyses are scheduled so that they are equally spaced between times Δ_t and t_{max} following pattern (2.3) rather than over the interval $(\Delta_{1,t}, t_{max})$. At interim analysis k, long-term responses $Y_{A,i}$ and $Y_{B,i}$, $i = 1, \ldots, n_{2,k}/2$, are available on treatments A and B, respectively. Short-term responses $X_{A,i}$ and $X_{B,i}$, $i = 1, \ldots, (n_{2,k} + n_{1,k})/2$, are also available, where

$$\left\{ X_{A,i}, X_{B,i} : i = \frac{n_{2,k}}{2} + 1, \ldots, \frac{n_{2,k} + n_{1,k}}{2} \right\}$$

are the measurements on the short-term endpoint for those partially observed subjects whose long-term response is not yet available. Let $\mathbf{X}^{(k)}$ denote the vector of all responses available at interim analysis k. Denote the design and variance-covariance matrices for the data available by $\mathbf{D}^{(k)}$ and $\mathbf{\Sigma}^{(k)}$ respectively, so that the data follow the model

$$\mathbf{X}^{(k)} \sim N(\mathbf{D}^{(k)}\boldsymbol{\beta}, \mathbf{\Sigma}^{(k)}).$$

The maximum likelihood estimator for $\boldsymbol{\beta}$ at interim analysis k, $\hat{\boldsymbol{\beta}}_k$, is given by

$$\hat{\boldsymbol{\beta}}_k = (\mathbf{D}^{(k)T}\mathbf{\Sigma}^{(k)-1}\mathbf{D}^{(k)})^{-1}\mathbf{D}^{(k)T}\mathbf{\Sigma}^{(k)-1}\mathbf{X}^{(k)}.$$

Let $\mathbf{c}_1 = (1, 0, -1, 0)^T$. Then, the maximum likelihood estimator for $\mu_{A,1} - \mu_{B,1}$ at interim analysis k can be extracted directly from $\hat{\boldsymbol{\beta}}_k$ using $\mathbf{c}_1^T\hat{\boldsymbol{\beta}}_k$, to obtain

$$\hat{\mu}_{A,1} - \hat{\mu}_{B,1} = \frac{2}{n_{2,k} + n_{1,k}} \sum_{i=1}^{(n_{2,k}+n_{1,k})/2} (X_{A,i} - X_{B,i}).$$

Similarly, defining $\mathbf{c}_2 = (0, 1, 0, -1)^T$, the maximum likelihood estimator for θ at

interim analysis k, $\hat{\theta}_k$, can be directly extracted using $\mathbf{c}_2^T\hat{\boldsymbol{\beta}}_k$ to obtain

$$\hat{\theta}_k = \frac{2}{n_{2,k}}\sum_{i=1}^{n_{2,k}/2}(Y_{A,i} - Y_{B,i}) - \frac{\tau\sigma_2}{\sigma_1}\left[\frac{2}{n_{2,k}}\sum_{i=1}^{n_{2,k}/2}(X_{A,i} - X_{B,i}) - (\hat{\mu}_{A,1} - \hat{\mu}_{B,1})\right]. \quad (7.1)$$

To gain more of an insight into the role measurements on the short-term endpoint play, it is helpful to first consider the scenario where we have $n_{2,k}$ fully observed subjects and $\mu_{A,1}$ and $\mu_{B,1}$ are known. Then, first considering responses on treatment A, it is helpful to write our model for the data as

$$\begin{aligned} X_{A,i} &= \mu_{A,1} + \epsilon_i, \\ Y_{A,i} &= \mu_{A,2} + \frac{\tau\sigma_2}{\sigma_1}\epsilon_i + \xi_i, \end{aligned} \quad (7.2)$$

for $i = 1, \ldots, n_{2,k}/2$, where error terms $\epsilon_i \sim N(0, \sigma_1^2)$, $\xi_i \sim N(0, \sigma_2^2(1-\tau^2))$ are independent. One can see that both responses share a common error term; a subject's short-term response allows us to observe this error term and so tells us something about $Y_{A,i}$. We can substitute $X_{A,i} - \mu_{A,1}$ for ϵ_i in (7.2) to help us estimate $\mu_{A,2}$; the mle for $\mu_{A,2}$ is

$$\hat{\mu}_{A,2} = \frac{2}{n_{2,k}}\sum_{i=1}^{n_{2,k}/2} Y_{A,i} - \frac{\tau\sigma_2}{\sigma_1}\left[\frac{2}{n_{2,k}}\sum_{i=1}^{n_{2,k}/2}(X_{A,i} - \mu_{A,1})\right]$$

If $\mu_{A,1}$ is unknown, each ϵ_i can be estimated by $X_{A,i} - \hat{\mu}_{A,1}$. If there are also $n_{1,k}/2$ subjects who are partially observed, then our estimate of $\mu_{A,1}$ is based on all $(n_{2,k} + n_{1,k})/2$ available short-term responses. Repeating this argument for $\mu_{B,2}$, we obtain $\hat{\theta}_k$ in (7.1). We conclude that the role of the short-term responses is quite subtle; it is not simply the case that as $|\tau|$ approaches 1, the short-term responses of partially observed subjects are substituted for their unobserved long-term responses.

If there are no partially observed subjects at an interim analysis, i.e., $\Delta_{1,t} = \Delta_t$, then $n_{1,k} = 0$ and $\hat{\theta}_k$ in (7.1) returns to $\bar{Y}_A - \bar{Y}_B$. For each $k = 1, \ldots, K-1$, let $Z_k = \hat{\theta}_k\sqrt{I_k}$ denote our standardised test statistic at interim analysis k for testing H_0. Define $\tilde{Z}_k$ to the standardised test statistic at decision analysis k. If termination is triggered at interim analysis k, we wait for the long-term responses of all $\tilde{n}_k$ subjects currently recruited to become available. Then,

$$\tilde{Z}_k = \frac{1}{\sigma_2\sqrt{\tilde{n}_k}}\sum_{i=1}^{\tilde{n}_k/2}(Y_{A,i} - Y_{B,i}),$$

as would be the case if no measurements on the short-term endpoint had been made.

Working from (7.1), the variance of the sampling distribution of $\hat{\theta}_k$ can be found in the usual way and shown to be given by

$$Var(\hat{\theta}_k) = \frac{4\sigma_2^2}{n_{2,k}} \left(1 - \tau^2 \frac{n_{1,k}}{(n_{1,k} + n_{2,k})}\right).$$

Galbraith & Marschner (2003) show this result holds asymptotically when σ_1^2, σ_2^2 and τ are unknown and mles are substituted in (7.1). If the short and long-term responses are indeed correlated, i.e., $\tau \neq 0$, the short-term data help us to make more precise estimates of θ. Under the proposed scheduling of the interim analyses, the information for θ accumulated at interim and decision analysis k is given by

$$I_k = \frac{(n_{1,k} + n_{2,k})}{4\sigma_2^2} \left\{1 + \frac{n_{1,k}}{n_{2,k}}(1 - \tau^2)\right\}^{-1}, \qquad \tilde{I}_k = \frac{\tilde{n}_k}{4\sigma_2^2}. \tag{7.3}$$

Suppose $\kappa = 0$, so that the secondary endpoint can be measured immediately upon commencement of treatment. Then, $\tilde{n}_k = n_{1,k} + n_{2,k}$ and the information, I_k, accrued if all recruited subjects were fully observed is $(n_{1,k} + n_{2,k})/4\sigma_2^2$. The factor

$$\left\{1 + \frac{n_{1,k}}{n_{2,k}}(1 - \tau^2)\right\}^{-1} \tag{7.4}$$

is a penalty for having $(\tilde{n}_k - n_{2,k})$ subjects in the pipeline. If $\tau = 0$, observations on the secondary endpoint provide no additional information for θ, i.e.,

$$I_k = \frac{(n_{1,k} + n_{2,k})}{4\sigma_2^2} \left\{1 + \frac{n_{1,k}}{n_{2,k}}\right\}^{-1},$$

as would be observed if no short-term endpoint was measured. If $\tau = 0.7$, $(1 - \tau^2)$ is 0.51 and we have approximately halved the penalty factor (7.4) and the damage done by having subjects in the pipeline at an interim analysis. Table 7.1 lists values of F_2 achieved by optimal two-stage GSTs for various values of τ when $r = 0.2$ and $\kappa = 0$. If no measurements on a short-term endpoint are made, the minima of $F_2 = 86.5\%$ of n_{fix}, an increase of 12% on n_{fix} compared to if response was immediate. However, if short-term measurements are made, with $\kappa = 0$ and $\tau = 0.7$, $F_2 = 83.2\%$. The term in $(1 - \tau^2)$ in (7.4) explains the sensitivity to τ when $\kappa = 0$ of the gains to be made by measuring a secondary endpoint.

For general $\kappa > 0$, some subjects will remain unobserved at the interim analysis. Then, in the limit as $|\tau| \to 1$

$$I_k = \frac{(n_{2,k} + n_{1,k})}{4\sigma_2^2} < \frac{\tilde{n}_k}{4\sigma_2^2},$$

so we can never hope to completely erase the damage done by having $(\tilde{n}_k - n_{2,k})$

			τ		
κ	0.5	0.6	0.7	0.8	0.9
0	85.0	84.2	83.2	81.9	80.2
0.2	85.2	84.6	83.8	82.7	81.4
0.4	85.5	85.0	84.4	83.6	82.6
0.6	85.8	85.4	85.0	84.5	83.9
1.0	86.5	86.5	86.5	86.5	86.5

Table 7.1: Values of F_2 achieved by optimal tests of $H_0 : \theta \leq 0$ when measurements on a short-term endpoint are available. Tests are designed and implemented with $\alpha = 0.05$, $\beta = 0.1$, $R = 1.15$ and $r = 0.2$.

subjects in the pipeline. Choosing our secondary endpoint so $n_{1,k}$ is large will increase the maximum benefits one can hope to make. In practice, there will be some trade off between ensuring high $|\tau|$ and $n_{1,k}$. Figure 7-2 illustrates how the information sequence for a two-stage delayed response GST when $r = 0.2$ will change when we measure a positively correlated secondary endpoint with $\tau = 0.9$ and $\kappa = 0.2$. Increasing the information at the interim analysis from I_1 to $I_1^{'}$ means that the minima of F_2 decreases from 86.5% of the fixed sample size to 81.4%; by making measurements on the short-term endpoint we have saved an additional 5% on n_{fix}. Table 7.1 lists minima of F_2 achieved by optimal two-stage tests when $r = 0.2$, for several values of τ and κ.

In the next section, we derive optimal delayed response GSTs to assess the efficiency gains that can be made by measuring a secondary endpoint.

7.2 Optimal tests

Figures 7-3(a)- 7-3(d) plot the maximum savings we can expect to make on the fixed sample test when data on a secondary endpoint are available. We consider properties of tests minimising F_2. We anticipate that these tests will perform well with respect to several alternative criteria; Eales & Jennison (1992) consider the case when response is immediate and $\alpha = \beta$ and find that GSTs minimising F_2 perform close to optimal with respect to other objective functions considered. Results for $\kappa = 1$ are included for reference and correspond to the case where no short-term endpoint is measured.

Our results underline the positive message that measuring a good secondary endpoint is a practical way in which we can recoup many of the benefits for early stopping that were associated with GSTs in the immediate response case. Even for larger values of r, the benefits of interim monitoring are much increased. We claim that group sequential

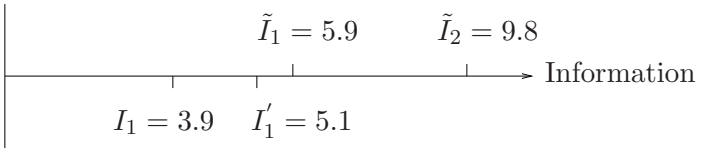
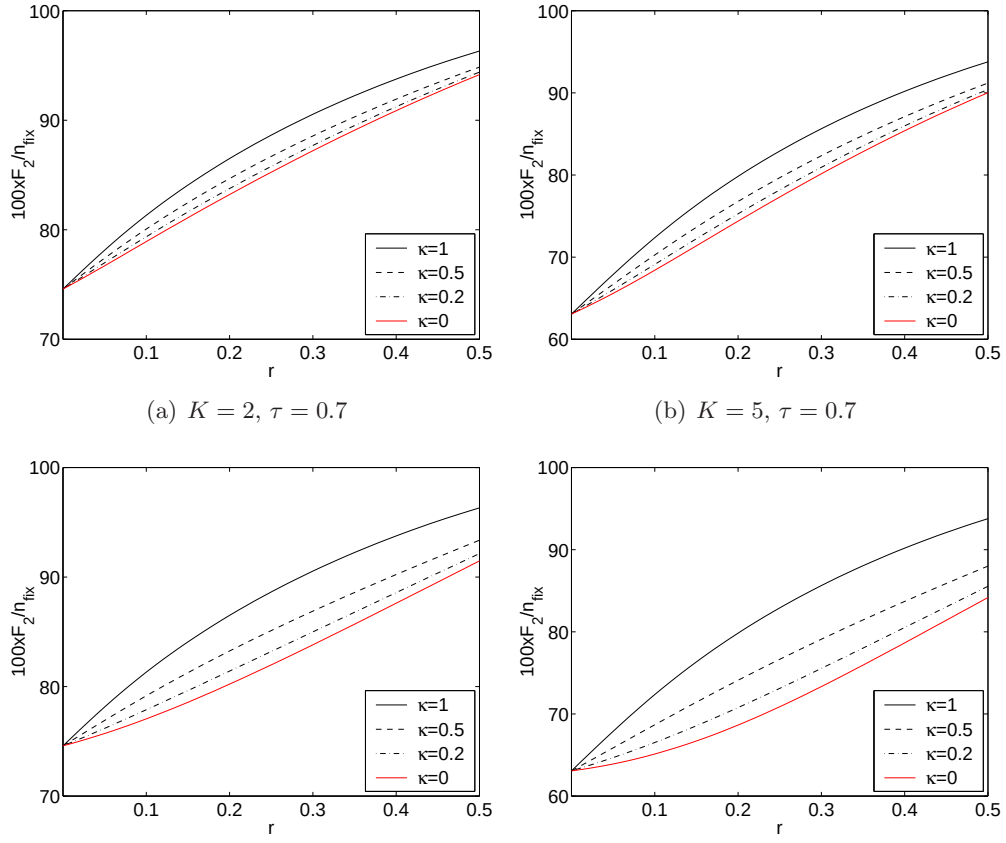


Figure 7-2: A two-stage delayed response GST with $\alpha = 0.05$, $\beta = 0.1$, $\delta = 1$, $R = 1.15$ and $r = 0.2$ will be based on information levels $\{I_1, \tilde{I}_1, \tilde{I}_2\}$. If data on a correlated secondary endpoint, with $\tau = 0.9$ and $\kappa = 0.2$, are available, these levels change to $\{I_1', \tilde{I}_1, \tilde{I}_2\}$.

testing now looks promising in some cases where we might previously have looked to carrying out a fixed sample test. For example, let $r = 0.3$ and $K = 5$. In the preamble to this chapter, we noted that delays of this magnitude are common in practice. If no short-term data are available, the expected sample size is 85.6% of the corresponding fixed sample size, representing a saving of more than 14%. However, if we can measure a secondary endpoint with $\kappa = 0.2$ and $\tau = 0.7$, our savings on n_{fix} are 19.1% which increase further to 24.5% when τ is as high as 0.9.

The sensitivity of the performance of our tests on τ is highlighted by our results; we do require high $|\tau|$ to take full advantage of the potential gains that can be made. For example, suppose we conduct a two-stage test with $r = 0.2$ and $\kappa = 0.2$. As τ increases from 0.5 to 0.6, our savings on the fixed sample test increase slowly from 14.8% to 15.4%. However, we see a more rapid increase from 16.2% to 17.3% as τ increases from 0.7 to 0.8.

Usually in practice, when choosing a secondary endpoint it will not be possible to simultaneously ensure both high τ and low κ. These competing aims pull in different directions since timing the secondary endpoint later may ensure stronger correlation with the long-term response but it will also mean that short-term data will be available for fewer pipeline subjects. From Figure 7-3, we see that the effects of this trade-off on our expected savings are quite complex. In some cases, timing our secondary endpoint later might be more efficient if this will effect a big increase in τ. For example, let $K = 5$ and $r = 0.2$. Suppose we can choose between two secondary endpoints. The first is such that $\kappa = 0.2$ and $\tau = 0.6$ and the second corresponds to $\kappa = 0.5$ and $\tau = 0.9$. In this case, opting for the second choice and waiting longer to observe the secondary endpoint increases our savings by almost 20%. However, ensuring high τ should not be our priority in every instance. For instance, if in our example τ only increased from 0.6 to 0.7, we would do better to opt for the first choice; opting for the second would actually result in a small decrease in expected savings. In many practical applications of course, we may not have the luxury of choice; there may not be an alternative secondary endpoint or we may be limited by logistical constraints. If we do



(a) $K = 2$, $\tau = 0.7$ (b) $K = 5$, $\tau = 0.7$

(c) $K = 2$, $\tau = 0.9$ (d) $K = 5$, $\tau = 0.9$

Figure 7-3: Minima of objective function F_2 expressed as a percentage of the corresponding fixed sample size attained by GSTs when measurements on a secondary endpoint are available. We consider tests with $\alpha = 0.05$, $\beta = 0.1$ and $R = 1.15$. Curves appear in the order they are listed in the legend.

have a choice however, a delicate balance must be struck between the competing aims of ensuring high τ and low κ so that we maximise the gains that can be made.

Our approach to dealing with data on a secondary endpoint assumes that our GST is to be based on a sequence of maximum likelihood estimates of a linear combination of parameters in a normal linear model. Hence, our approach extends naturally to the case where repeated measurements are made on each subject if data can be represented as a normal linear model. This scenario is likely to occur often in practice. For example, Grieve & Krams (2005) discuss the ASTIN trial into stroke where a longitudinal model is used to predict a subject's Scandinavian Stroke Scale (SSS) score at 12 weeks using earlier scores at 1, 4 and 8 weeks known to be correlated with the primary endpoint. In the next sections, we continue with the problem of formulating tests of H_0 when observations are made on a short-term and long-term endpoint but extend our methods to the case when the covariance matrix of the joint model for the data is unknown.

7.3 Dealing with unknown nuisance parameters

7.3.1 Introduction

In the previous sections, methodology and results were presented for the case where σ_1^2, σ_2^2 and τ are known. Computations requiring knowledge of these nuisance parameters, such as calculating the target sample size and also, at each analysis, the observed information levels and mles of θ, can be done exactly. In general, however, σ_1^2, σ_2^2 and τ will not be known and methodology for dealing with this case is of substantial practical interest.

Suppose we measure a single long-term endpoint where response is immediate and data are normally distributed with unknown variance σ^2. We estimate n_{max} given an initial guess at σ^2; this target is then updated as a more accurate estimate of σ^2 becomes available. Internal pilot studies stipulate that an estimate of σ^2 based on data accumulated in the first part of the study is substituted into some pre-specified rule to set the sample size for the next stage. At the end of the study, the test of H_0 uses the estimate of σ^2 based on accumulated data from both stages. Numerous variants on this strategy have been proposed, for example by Wittes & Brittain (1990) and Gould & Shih (1992) amongst others. Wittes & Brittain note that the final estimate of σ^2 in an internal pilot study will be biased downwards. This is because overestimates at the end of the first stage will be diluted by larger second stage sample sizes than underestimates. This bias leads to an inflation of the type I error probability, although Jennison & Turnbull (2000, Chapter 14) show this to be small so long as the sample size in the internal pilot is sufficiently large. Keiser & Friede (2000) propose an internal pilot study which controls the type I error rate exactly.

One can also conduct sample size re-estimation within a group sequential framework. Denne & Jennison (1999) derive group sequential t-tests, where updated estimates of σ^2 are used to adjust the sample size according to some pre-specified rule. However, these tests do not control the type I error rate at exactly level α. Extending the conditional rejection principle of Müller & Schäfer (2001), Timmesfeld et al. (2007) derive an exact group sequential t-test where sample size adjustments do not have to follow a pre-specified rule.

7.3.2 Information based interim monitoring

Mehta & Tsiatis (2001) propose an "information monitoring approach" as a flexible alternative to group sequential t-tests. Their approach stipulates that maximum

information error spending designs be used in conjunction with sample size re-estimation. Suppose we observe responses $X_{A,i} \sim N(\mu_A, \sigma^2)$ and $X_{B,i} \sim N(\mu_B, \sigma^2)$ on treatments A and B and wish to test $H_0 : \theta = \mu_A - \mu_B \leq 0$ against $H_1 : \theta > 0$. We proceed according to the following algorithm:

1. Select a pair of error spending functions, $f(t)$ and $g(t)$, which will be used to construct the boundaries of our GST. Then, assuming a certain information sequence, for example $I_k = kI_{max}/K$, for $k = 1, \ldots, K$, we can calculate the maximum information level I_{max} required for the error spending test to terminate properly at stage K with $l_K = u_K$. An initial guess at the nuisance parameter σ^2 can then be used to make a preliminary estimate of the maximum sample size needed to reach this target.

2. To implement the test, for each $k = 1, 2, \ldots$, let s_k^2 denote the usual unbiased estimate of σ^2 at interim analysis k which follows marginal distribution $\sigma^2 \chi^2_{n_k-2}/(n_k - 2)$. Substituting this estimate for σ^2, we calculate an estimate of our current information for θ, denoted $\hat{I}_k(s_k^2)$, and obtain the fraction of the target information level that has been accumulated as $\hat{t}_k = \hat{I}_k(s_k^2)/I_{max}$. If n_k subjects have been observed, divided equally between both treatments, the t-statistic for testing $H_0 : \theta = 0$ at interim analysis k as

$$T_k = \frac{\bar{X}_{A,k} - \bar{X}_{B,k}}{\sqrt{s_k^2(4/n_k)}},$$

   where $T_k \sim t_{n_k-2}$ under $\theta = 0$.

3. By interim analysis k, we have observed information levels $\hat{I}_1(s_1^2), \ldots, \hat{I}_k(s_k^2)$. Given this sequence, boundary constants l_k and u_k are calculated for monitoring a sequence of standardised test statistics with canonical distribution (2.5), i.e., as for normal data with known variance. If $\hat{I}_k(s_k^2) < I_{max}$, the type I and type II error probabilities, $\pi_{1,k}$ and $\pi_{2,k}$, to be spent at stage k are

$$\begin{aligned} &\pi_{1,1} = f(\hat{t}_1) && \pi_{2,1} = g(\hat{t}_1) \\ &\pi_{1,k} = f(\hat{t}_k) - f(\hat{t}_{k-1}) && \pi_{2,k} = g(\hat{t}_k) - g(\hat{t}_{k-1}) \qquad k = 2, 3, \ldots. \end{aligned}$$

   If $\hat{I}_k(s_k^2) \geq I_{max}$, u_k is calculated so that we spend all the remaining type I error probability and we set $l_k = u_k$ to ensure the test terminates properly. Constants for monitoring the t-statistics are computed so that we preserve the marginal

probabilities of stopping at stage k under $\theta = 0$. This gives the stopping rule,

$$\begin{aligned} &\text{if } T_k \geq t_{n_k-2,1-\Phi(u_k)} && \text{stop, reject } H_0, \\ &\text{if } T_k \leq t_{n_k-2,1-\Phi(l_k)} && \text{stop, accept } H_0, \\ &\text{otherwise} && \text{continue to stage } k+1, \end{aligned}$$

where $t_{\nu,q}$ denotes the upper q tail point of a t-distribution on ν degrees of freedom.

4. If we decide to continue, then given s_k^2 we re-estimate the sample size required to reach the target information level I_{max} and choose the next group size accordingly.

When monitoring information, adjusting the sample size in response to updated nuisance parameter estimates seems natural as we work towards attaining a target information level. The justification for monitoring a sequence of t-statistics using error spending boundaries designed for normal data with known variance relies on large sample results. However, simulation studies conducted by Mehta & Tsiatis (2001) and Jennison & Turnbull (2007) have shown that in most cases type I error rates are controlled at levels close to their nominal values in small samples.

In the next section, we extend the information based monitoring approach of Mehta & Tsiatis to the case where data are collected on a primary and secondary endpoint with delays Δ_t and $\Delta_{1,t}$, respectively, and the covariance matrix of their joint model is unknown.

7.4 Information based monitoring for delayed responses

7.4.1 Introduction

Recall the testing problem outlined in Section 7.1. We wish to test group sequentially our null hypothesis $H_0 : \theta \leq 0$ against the one-sided alternative $H_1 : \theta > 0$ in a maximum of K stages, with type I error rate α at $\theta = 0$ and power $1 - \beta$ at $\theta = \delta$. In this section, we explain how, using the error spending tests for delayed responses developed in Chapter 6, it is possible to combine group sequential testing with sample size re-estimation using maximum information error spending tests. Prior to the start of the trial, we choose two error spending functions, f and g, which will govern how type I and II error probabilities are spent during the course of the trial. In the work presented here, functions are selected from the ρ family, indexed by the parameter $\rho > 0$. We may wish to choose ρ so that error probabilities are spent slowly in the early

stages when our nuisance parameter estimates are based on lower degrees of freedom.

For design purposes, we make some assumptions about likely values of the nuisance parameters. Let $\sigma_{2,0}$ and τ_0 denote initial guesses at the nuisance parameters σ_2 and τ. Fixing the values of α, β and δ, the fixed sample test of H_0 with the required error probabilities requires information

$$I_{fix} = \frac{\{\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)\}^2}{\delta^2}$$

and we estimate that it requires a total sample size $n_{fix,0} = 4\sigma_{2,0}^2 I_{fix}$, with patient entry divided equally between treatments A and B. Assuming accrual proceeds at constant rate c, recruitment is completed in time $t_{fix,0} = n_{fix,0}/c$.

We fix our choice of ρ and hence the error spending functions f and g we will work with. We design our GST anticipating a maximum number of analyses K. Delays Δ_t and $\Delta_{1,t}$ in the long and short-term endpoints are also fixed which determines $\kappa = \Delta_{1,t}/\Delta_t$. Define $\lambda_0 = \Delta_t/t_{fix,0}$. Given these values, we conduct a bisection search over values of R to find the target information level $I_{max} = RI_{fix}$ required for a K-stage delayed response error spending test of H_0 with type I error rate α at $\theta = 0$ and power $1-\beta$ at $\theta = \delta$ to terminate properly with $l_K = u_K$. For a given value of R, our preliminary estimate of the required total maximum total sample size is $n_{max,0} = 4\sigma_{2,0}^2 RI_{fix}$. Although the test will be implemented using the information levels that are actually observed, we assume for design purposes that accrual proceeds at a constant rate c so that at interim analysis $k = 1, \ldots, K-1$,

$$n_{2,k} = \frac{n_{max,0}}{K}[k(1-\lambda_0/R)], \quad \tilde{n}_k = \frac{n_{max,0}}{K}\left[\frac{k(R-\lambda_0) + K\lambda_0}{R}\right]$$

and $n_{1,k}$ remains constant at $n_{fix,0}\lambda_0(1-f)$, with all numbers equally divided between treatments A and B. When searching over values of R, we know that if it is too high the boundaries will terminate with $l_k > u_k$ at stage $k < K$, while if R is too low, we will terminate with $l_K < u_K$. Once the necessary inflation factor has been found, the target information level which we will aim to reach upon the test's implementation is set at $I_{max} = RI_{fix}$. Hence, our preliminary estimate of the test's maximum total sample size is $n_{max,0} = Rn_{fix,0}$. When determining the test's first group size, we assume accrual will proceed at the same constant rate c as was anticipated at the design stage. Let $t_{max,0}$ denote the time taken to recruit $n_{max,0}$ subjects at constant accrual rate c. Then, the first group size is calculated anticipating $K-1$ interim analyses equally spaced between times Δ_t and $t_{max,0}$. Hence, at the first interim analysis we have

$$n_{2,1} = \frac{1}{K}\, n_{max,0}(1-\lambda_0/R), \qquad \tilde{n}_1 = \frac{n_{max,0}}{K}\left[\frac{R + \lambda_0(K-1)}{R}\right]$$

and $n_{1,1} = n_{fix,0}\lambda_0(1-f)$, with all numbers equally divided between treatments A and B.

In the next sections, we explain how to implement our proposed information based monitoring approach for delayed responses. The algorithm we follow adapts that of the Mehta & Tsiatis (2001) approach for testing H_0 with the maximum information error spending designs formulated in Chapter 6.

7.4.2 Implementation

We begin by explaining how to update estimates of our nuisance parameters at each stage. At each interim analysis k, we assume that there are equal numbers of subjects recruited, fully and partially observed on each treatment. In order to calculate our estimate of σ_1^2 at an interim analysis, we naturally use the short-term observations available on both the fully and partially observed subjects. Denote the usual unbiased pooled estimates of σ_1^2 by $s_{1,k}^2$, $k = 1, 2, \ldots$, so each $s_{1,k}^2 \sim \sigma_1^2 \chi_{m_k}^2/m_k$, where $m_k = n_{1,k} + n_{2,k} - 2$. Likewise, denote the unbiased pooled estimates of σ_2^2 by $s_{2,k}^2$, $k = 1, 2, \ldots$, so $s_{2,k}^2 \sim \sigma_2^2 \chi_{\nu_k}^2/\nu_k$, where $\nu_k = n_{2,k} - 2$. Let $\bar{x}_{A,k}$ denote the sample mean of the $(n_{2,k} + n_{1,k})/2$ short-term observations accrued on treatment A and define $\bar{x}_{B,k}$ similarly. At interim analysis $k = 1, 2 \ldots$, we estimate τ by

$$\hat{\tau}_k = \frac{\sum_{i=1}^{n_{2,k}/2}(x_{A,i} - \bar{x}_{A,k})(y_{A,i} - \bar{y}_{A,k})}{(n_{2,k}/2 - 1)s_{1,k}\, s_{2,k}}.$$

At a decision analysis, it turns out we need only estimate σ_2^2. Let $\tilde{s}_{2,k}^2$, $k = 1, \ldots, K$, denote the unbiased pooled estimate of σ_2^2 calculated based on $\tilde{n}_k$ long-term observations, so each $\tilde{s}_{2,k}^2 \sim \sigma_2^2 \chi_{\tilde{n}_k-2}^2/(\tilde{n}_k - 2)$. Let $I_k(s_{2,k}, \hat{\tau}_k)$, $k = 1, 2, \ldots$, denote an estimate of the observed information at interim analysis k computed by substituting $s_{2,k}^2$ and $\hat{\tau}_k$ into formula (7.3). Let $\tilde{I}_k(\tilde{s}_{2,k})$, $k = 1, 2, \ldots$, denote the information at decision analysis k as estimated using $\tilde{s}_{2,k}^2$.

In order to implement our test, we must define test statistics at the interim and decision analyses. At interim analysis $k = 1, 2, \ldots$, let $\hat{\theta}_k(s_{1,k}, s_{2,k}, \hat{\tau}_k)$ be our estimate of θ calculated by substituting current estimates for the nuisance parameters into our formula for $\hat{\theta}_k$. We define the test statistic based on this quantity to be

$$T_k = \hat{\theta}_k(s_{1,k}, s_{2,k}, \hat{\tau}_k)\sqrt{I_k(s_{2,k}, \hat{\tau}_k)}.$$

Since $\hat{\theta}_k(s_{1,k}, s_{2,k}, \hat{\tau}_k)$ and $I_k(s_{2,k}, \hat{\tau}_k)$ depend upon $s_{1,k}^2$, $s_{2,k}^2$ and $\hat{\tau}_k$ in a complicated way, the test statistic T_k does not follow a standard distribution. At decision analysis

$k = 1, 2, \ldots$, our mle of θ and $\tilde{I}_k$ depend only upon the unknown σ_2^2. Our t-statistic for testing H_0 is given by

$$\tilde{T}_k = \frac{\sum_{i=1}^{\tilde{n}_k/2} Y_{A,i} - \sum_{i=1}^{\tilde{n}_k/2} Y_{B,i}}{\sqrt{\tilde{s}_{2,k}^2 \, \tilde{n}_k}},$$

which follows a $t_{\tilde{n}_k - 2}$ distribution under $\theta = 0$.

We propose that error probabilities are spent as a function of $\tilde{I}_k$ rather than I_k. Adopting this approach here, we calculate a preliminary estimate of $\tilde{I}_k$ at interim analysis k using $s_{2,k}^2$ and plug this into our error spending functions f and g. It is possible that we may observe either $I_k(s_{2,k}, \hat{\tau}_k) < I_{k-1}(s_{2,k-1}, \hat{\tau}_{k-1})$ or $\tilde{I}_k(s_{2,k}) < \tilde{I}_{k-1}(s_{2,k-1})$, due to changes in the estimates of the nuisance parameters as more data accumulate. Jennison (2006) comments that when monitoring information in an immediate response setting, decreasing sequences of information occur surprisingly often, particularly when estimates are based on low degrees of freedom. In this situation, we choose to follow his pragmatic solution and do not allow early stopping. Otherwise, if $k = 1$, we spend type I and type II error probabilities

$$\pi_{1,1} = f(\tilde{I}_1(s_{2,1})/I_{max})$$
$$\pi_{2,1} = g(\tilde{I}_1(s_{2,1})/I_{max}).$$

and for $k = 2, \ldots$, we spend

$$\pi_{1,k} = f(\tilde{I}_k(s_{2,k})/I_{max}) - f(\tilde{I}_{k-1}(s_{2,k-1})/I_{max})$$
$$\pi_{2,k} = g(\tilde{I}_k(s_{2,k})/I_{max}) - g(\tilde{I}_{k-1}(s_{2,k-1})/I_{max}).$$

We approximate by computing boundaries for monitoring a sequence of standardised statistics $\{Z_k, \tilde{Z}_k\}$ following the usual canonical distribution. Then, using the strategy outlined in Chapter 6, we find the boundary constants l_k, u_k and c_k so that given information levels $I_1(s_{2,1}, \hat{\tau}_1), \ldots, I_k(s_{2,k}, \hat{\tau}_k)$, $\tilde{I}_k(s_{2,k})$,

$$\pi_{1,k} = \psi_k(l_1, u_1, \ldots, l_{k-1}, u_{k-1}, l_k, u_k, c_k; \theta = 0)$$
$$\pi_{2,k} = \xi_k(l_1, u_1, \ldots, l_{k-1}, u_{k-1}, l_k, u_k, c_k; \theta = \delta),$$

where ψ_k and η_k are as defined in (2.6)-(2.9) of Chapter 2. Our approximating sequence of statistics are correlated, with

$$corr(Z_{k-1}, Z_k) = \sqrt{I_{k-1}/I_k}$$

and

$$corr(Z_k, \tilde{Z}_k) = \sqrt{I_k/\tilde{I}_k}.$$

We mis-specify these correlations because of their dependence on τ, which has to be estimated. However, referring to (7.3), we see that the information ratios I_{k-1}/I_k and $I_k/\tilde{I}_k$ do not depend on σ_2^2, since this unknown cancels out. Hence, when calculating these ratios, rather than plugging in the estimated information levels directly we increase our accuracy by using the formulae

$$\frac{I_{k-1}}{I_k} = \frac{n_{2,k-1}(n_{1,k-1}+n_{2,k-1})}{n_{2,k}(n_{1,k}+n_{2,k})} \left\{ \frac{(1-\hat{\tau}_k^2)n_{1,k}+n_{2,k}}{(1-\hat{\tau}_{k-1}^2)n_{1,k-1}+n_{2,k-1}} \right\}.$$

The ratios $I_k/\tilde{I}_k$ are calculated similarly.

Suppose $\tilde{I}_k(s_{2,k}) < I_{max}$ at interim analysis k. In this case we must decide at the interim analysis whether or not to close recruitment. Our boundary constants l_k, u_k and c_k are set for monitoring a sequence of Z-statistics. Rather than comparing the T_k directly against these values, we could adopt a "significance level" approach. This entails setting critical values for the T_k which preserve the marginal probabilities under $\theta = 0$ that $Z_k \geq u_k$ and $Z_k \leq l_k$. When this approach is used to adapt GSTs designed for normal data with known variance to the case when the variance is unknown, Jennison & Turnbull (2001) show that error rates are controlled at levels close to their nominal values. When working with the T_k, things are rather more complicated because these test statistics do not follow a standard distribution although one could use simulation to generate their empirical distribution under $\sigma_1^2 = s_{1,k}^2$, $\sigma_2^2 = s_{2,k}^2$ and $\tau = \hat{\tau}_k$. We can then compare T_k with certain percentiles of this distribution, closing recruitment if T_k either exceeds the upper $1-\Phi(u_k)$ percentile or is less than the $1-\Phi(l_k)$ percentile. Whilst one might pursue this approach in practice, for the purposes of our simulations we adopt the cruder but computationally less intensive approach of treating the T_k as standard normal variates under H_0 and comparing directly with l_k and u_k.

Suppose recruitment is closed at the interim analysis and we continue to the decision analysis. Marginally, $\tilde{T}_k$ has a $t_{\tilde{n}_k-2}$ distribution under $\theta = 0$. Our estimate of $\tilde{I}_k$ at this stage, based on $\tilde{s}_{2,k}^2$, is likely to differ from our preliminary estimate of this quantity made at the interim analysis. Despite this, we have chosen to keep our decision constant fixed at c_k, even if $\tilde{I}_k(\tilde{s}_{2,k}) < I_k(s_{2,k}, \hat{\tau}_k)$. Applying the significance level approach at decision analysis k, we reject H_0 if

$$\tilde{T}_k \geq t_{\tilde{n}_k-2,1-\Phi(c_k)}, \qquad k = 1, 2, \ldots. \tag{7.5}$$

If either $\tilde{I}_k(s_{2,k}) \geq I_{max}$ at an interim analysis or $k = K$, recruitment is automatically closed; we wait for all subjects in the pipeline to be fully observed before conducting a final analysis. In the notation of Chapter 6, u_k is calculated so that all remaining

type I error probability is spent under information level $\tilde{I}_k(\tilde{s}_{2,k})$, and we set $l_k = u_k$ to ensure the test terminates properly. Upon reaching the decision analysis, we may observe $\tilde{I}_k(\tilde{s}_{2,k}) < I_{max}$. However, recruitment cannot be reopened and setting $u_k = c_k$ in decision rule (7.5), we decide whether to reject or accept H_0 accordingly.

Suppose we decide to continue recruitment at an interim analysis. Using our current estimate of σ_2^2, our updated estimate of the required sample size is given by

$$n_{max,k} = 4s_{2,k}^2 I_{max}.$$

Assuming recruitment continues at the same constant rate c as was anticipated at the design stage, our next interim analysis is timed so that the remaining $(n_{max,k} - \tilde{n}_k)$ subjects yet to be recruited will be accrued in $(K - k)$ equally sized groups. For each $k = 1, 2, \ldots$, at interim analysis $(k + 1)$ we have

$$\tilde{n}_{k+1} = \tilde{n}_k + \frac{n_{max,k} - \tilde{n}_k}{K - k} \qquad n_{2,k+1} = \frac{n_{max,k} + \tilde{n}_k[K - k - 1]}{K - k} - \lambda_0 n_{fix,0}.$$

The number of partially observed subjects at interim analysis $(k + 1)$ will be $n_{fix,0}\lambda_0(1 - f)$.

7.4.3 Simulation results

In this section, we verify through simulation that the information based monitoring scheme discussed in Sections 7.4.1 and 7.4.2 does indeed control the type I error rate and maintain the power of a test at close to their nominal values. Tests are designed and implemented for $\kappa = \Delta_{1,t}/\Delta_t = 0.6$. We look at maximum information trials designed under initial guesses $\sigma_{1,0}^2 = 0.8$, $\sigma_{2,0}^2 = 2.3$ and $\tau_0 = 0.6$ or 0.8 to have type I error probability $\alpha = 0.05$ at $\theta = 0$, power 0.9 at $\theta = \delta$ when $K = 5$. We present results for the case where error probabilities are spent according to functions

$$f(t) = \alpha \min\{t^2, 1\} \qquad g(t) = \beta \min\{t^2, 1\} \quad \text{for } t \geq 0.$$

Upon implementation, data are distributed with $\sigma_1^2 = 1$, $\sigma_2^2 = 2$ and $\tau = 0.7$. Tests follow the algorithm outlined in Section 7.4.2, assuming accrual proceeds at the same constant rate c that was anticipated at the design stage. A maximum of $K = 5$ analyses are allowed; if the target information level has not been reached by the fifth analysis, all of the remaining type I error probability is spent and we set $u_5 = l_5$ to ensure termination. Type I error rates and power are estimated via simulation using $50,000$ replicates; standard errors are 0.001 for type I error probabilities and 0.0013 for power. Table 7.2 lists results for several combinations of δ, the alternative at which we specify power, τ_0, our initial guess at the correlation coefficient and $\lambda_0 = \Delta_t/t_{fix,0}$.

		$\tau_0 = 0.6$			$\tau_0 = 0.8$		
	λ_0	$n_{max,0}$	Type I error rate	Power	$n_{max,0}$	Type I error rate	Power
$\delta = 0.5$	0.1	361.5	0.052	0.893	360.0	0.051	0.894
	0.2	373.6	0.052	0.892	371.2	0.051	0.890
	0.3	382.9	0.051	0.891	380.3	0.051	0.888
$\delta = 1.0$	0.1	90.4	0.056	0.877	90.0	0.056	0.876
	0.2	93.4	0.053	0.869	92.8	0.055	0.868
	0.3	95.7	0.052	0.869	95.1	0.053	0.866

Table 7.2: Attained error rates of an information monitoring procedure for delayed responses. The table gives type I error probabilities and power attained by tests with a maximum of $K = 5$ stages designed to achieve type I error probability 0.05 at $\theta = 0$, power 0.9 at $\theta = \delta$ under delay parameter λ_0. Data are simulated under $\sigma_1^2 = 1$, $\sigma_2^2 = 2$ and $\tau = 0.7$ but the test is designed under initial guesses $\sigma_{1,0}^2 = 0.8$, $\sigma_{2,0}^2 = 2.3$ and τ_0. Tests are designed and implemented under $\kappa = 0.6$

In Table 7.2, the target information level for each test can be calculated as $I_{max} = n_{max,0}/4\sigma_{2,0}^2$. Our results show that when studies are implemented under smaller target information levels there are larger deviations in the attained error rates from their nominal values. For example, when $\delta = 1$ and $\tau_0 = 0.6$, $I_{max} = 9.8$. In this case, the attained type I error rate is 0.056 and power at $\theta = 1$ is 0.877. However, the method performs much better for smaller values of δ. This is to be expected since there are now a greater number of degrees of freedom with which to estimate the three unknown nuisance parameters. For $\delta = 0.5$, the method performs well and attained error rates are very close to their nominal values. For example, for $\lambda_0 = 0.1$ and $\tau_0 = 0.6$, $I_{max} = 39.3$ and the attained type I error rate is 0.052. The method's accuracy is also robust to the misspecification of τ at the design stage. Whether this error arises because we are too optimistic concerning our choice of secondary endpoint or too pessimistic appears to have little impact. We conclude that in the context of a Phase III trial, when fairly precise estimates of the nuisance parameters are available and smaller values of δ are of interest, one can have confidence that our information based monitoring approach controls the type I error rate at a level close to α. Drawing together these conclusions with those reached in early sections of this chapter, we see that we have formulated an efficient approach to dealing with delayed responses when measurements can be made on a suitable short-term endpoint which is flexible enough to be used in practice.

Chapter 8

Optimal delayed response tests for a combined objective

8.1 Introduction

Recall the testing scenario considered in earlier chapters of this thesis: subjects are randomised to either an experimental treatment or control and, after a delay of length Δ_t, measurements are made on a single endpoint of direct clinical interest. Thus, we have responses $X_{A,i} \sim N(\mu_A, \sigma^2)$, $i = 1, 2, \ldots$, for subjects allocated to the new treatment and $X_{B,i} \sim N(\mu_B, \sigma^2)$, $i = 1, 2, \ldots$, for those on control, where σ^2 is known. Define $\theta = \mu_A - \mu_B$. In Chapters 3 to 7, our priority was finding highly efficient tests of $H_0 : \theta \leq 0$ for delayed responses with low average expected sample sizes. However, minimising the number of recruited subjects may not always be our only concern. Often, we are also desirous of bringing the trial to a conclusion as quickly as possible. Promptly identifying a successful treatment means we can jump to the next stage of development, ultimately reducing the time taken for our drug to reach market. Meanwhile, if the treatment is ineffective, stopping early for futility enables us to focus our attention on planning trials of other compounds in earlier stages of development. The combined aims of low sample size and rapid time to a conclusion can pull in opposite directions when formulating group sequential tests of H_0. It has already been noted that when there is a delay in response, the benefits of slowing down recruitment for reducing a test's expected sample size must be balanced against the impact of this strategy on its overall time scale. With such trade-offs in mind, in this chapter we shall derive optimal tests which are efficient for both expected sample size and time to a conclusion and explore the gains to be made by testing group sequentially when there is a delay in response.

In previous chapters, GSTs for delayed responses have been formulated assuming one is obliged to follow-up those subjects in the pipeline when termination was triggered before deciding whether to reject or accept H_0. However, this will not always be pertinent. Reporting the overrun data is only appropriate if pipeline subjects continue to be treated according to protocol. For example, if those involved in the study are unblinded or subjects randomised to the inferior treatment switched, using these data to make inferences could bias our conclusions. Allowing immediate stopping at an interim analysis for rejection or acceptance of H_0 may also be apt in the context of a confirmatory Phase III trial. If the decision is for success, one can then start preparing to file for regulatory approval without delay, although we expect that the overrun data will still be reported informally at a later date. In this chapter, we shall derive optimal versions of GSTs which permit two types of stopping at an interim analysis: we close recruitment and make a decision either immediately or postpone doing so until all current pipeline information becomes available. If our sole objective is to minimise expected sample size, it is always optimal to stop and wait for the overrun data; under this model, there is a charge for recruiting subjects but no penalty for waiting to measure their response. However, if there is a shift in objective and we also seek a rapid time to a conclusion, it may well be optimal in some instances to terminate immediately rather than wait for the pipeline data.

When we are looking for GSTs which are optimal for a combined objective involving expected sample size and expected time to a conclusion, we expect to find K-stage tests of the form

At interim analysis $k = 1, \ldots, K$,

if $l_k < Z_k < u_k$	continue to interim analysis $k+1$,
if $l_k^\star < Z_k \leq l_k$ or $u_k \leq Z_k < u_k^\star$	terminate recruitment, proceed to decision analysis k,
if $Z_k \geq u_k^\star$	terminate immediately and reject H_0,
if $Z_k \leq l_k^\star$	terminate immediately and accept H_0.

At decision analysis $k = 1, \ldots, K$

if $\tilde{Z}_k \geq c_k$	reject H_0,	(8.1)
if $\tilde{Z}_k < c_k$	accept H_0.	

Patient entry into the study is balanced so that there are equal numbers on each treatment. Let n_{fix} denote the total number of subjects required by a fixed sample test of $H_0 : \theta \leq 0$ of size α to achieve power $1 - \beta$ at $\theta = \delta$. If accrual proceeds

at a constant rate of c subjects per unit of time, the fixed sample test comes to conclusion, i.e., recruitment is completed and a decision of whether to reject H_0 made, after $t_{fix} = \Delta_t + n_{fix}/c$ units of time have elapsed. Formulating a GST of $H_0 : \theta \leq 0$, we set the maximum sample size to be $n_{max} = Rn_{fix}$. Under an accrual rate of c, recruitment will be completed at time $t_{max} = n_{max}/c$. Define the delay parameter $r = \Delta_t/t_{max}$. Let t_k and $\tilde{t}_k$ denote the calendar timings of interim and decision analysis k when a total of n_k and $\tilde{n}_k$ responses are available, respectively. Scheduling analyses at calendar times

$$t_k = \frac{k}{K} t_{max}(1-r) + rt_{max}, \qquad \tilde{t}_k = t_k + rt_{max}, \qquad \text{for } k = 1, \ldots, K,$$

generates the observed information sequence

$$I_k = \frac{k}{K}(1-r)I_{max}, \qquad \tilde{I}_k = I_k + rI_{max}, \qquad \text{for } k = 1, \ldots, K.$$

Suppose there is an additional delay of ϵ incurred by waiting to clean and transfer data ready for an interim analysis. If recruitment continues during this period, by the time of the analysis there will be $c(\Delta_t + \epsilon)$ subjects in the pipeline; if termination is then triggered, we must wait a further $\Delta_t + \epsilon$ units of time before all the pipeline information is available and has been cleaned. Define $r' = (\Delta_t + \epsilon)/t_{max}$. Then, to take account of this overrun, the study would therefore be planned so that in preparation for interim analysis k we lock the analysis data set at time

$$t'_k = r't_{max} + \frac{k}{K}(1-r')t_{max} - \epsilon$$

when information $I_k = k(1-r')I_{max}/K$ has been observed. The data are cleaned and subsequently analysed at time $t_k = t'_k + \epsilon$; if a boundary is crossed, a decision analysis is conducted at time $\tilde{t}_k = t_k + r't_{max}$ and information level $I_k + r'I_{max}$. Hence the study can terminate at times

$$t_k = \frac{k}{K} t_{max}(1-r') + r't_{max}, \qquad \tilde{t}_k = t_k + r't_{max}, \qquad \text{for } k = 1, \ldots, K;$$

it is as if the study had been planned for a delay $\Delta_t + \epsilon$ in the primary endpoint and immediate data transfer. The same test will be optimal in both cases. Hence, from now on we make the simplifying assumption that data transfer is immediate, knowing that if it is not, one can read off the results for this case by looking for the appropriate value of r.

Define T to represent the time taken for the GST to reach a conclusion and N to represent the total number of subjects recruited on its termination. Note that we can choose to measure time in several different units. However, the underlying design

problem does not change if, for example, we shift from measuring calendar time in weeks to months. With this in mind, we concentrate on finding GSTs of $H_0 : \theta \leq 0$ for delayed responses minimising weighted averages of $\mathbb{E}_\theta(N)/n_{fix}$ and $\mathbb{E}_\theta(T)/t_{fix}$ under different values of θ; tests minimising averages of these ratios will be invariant to the units in which time is measured. Let a and b are non-negative constants satisfying $a + b = 1$. Then, generalising objective functions F_i, $i = 1, \ldots, 4$, defined in Section 3.1, our objective is to find K-stage GSTs tests minimising

$$G_1 = \mathbb{E}\left(\frac{aN}{n_{fix}} + \frac{bT}{t_{fix}}; \theta = \frac{\delta}{2}\right), \quad G_2 = \sum_{i=0}^{1} \mathbb{E}\left(\frac{aN}{n_{fix}} + \frac{bT}{t_{fix}}; \theta = i\delta\right)$$

$$G_3 = \sum_{i=0}^{1} \mathbb{E}\left(\frac{aN}{n_{fix}} + \frac{bT}{t_{fix}}; \theta = \frac{\delta(4i-1)}{2}\right),$$

$$G_4 = \int_\Theta \mathbb{E}\left(\frac{aN}{n_{fix}} + \frac{bT}{t_{fix}}; \theta\right) \frac{2}{\delta}\phi\left(\frac{\theta - \delta/2}{\delta/2}\right) d\theta,$$

with type I error rate α at $\theta = 0$ and power $1-\beta$ at $\theta = \delta$. Setting $b = 0$ corresponds to those problems solved in earlier chapters of this thesis, where for each $i = 1, \ldots, 4$, tests minimising F_i also minimise G_i. To find the frequentist tests optimal for other values of a and b, we adopt the approach taken in Chapter 3 and search for an appropriate unconstrained Bayes problem whose solution is the test we seek. We shall explain in greater detail how this can be done in the next few sections.

8.2 Finding optimal tests

8.2.1 Formulation of the problem to be solved

In this section we explain how to create a Bayes problem whose solution is the optimal GST minimising G_1 with type I error rate α at $\theta = 0$ and power $1-\beta$ at $\theta = \delta$. A Bayes problem whose solution minimises a weighted average of expected sample size and time to a conclusion charges for sampling both subjects and time under certain values of θ. Define sampling cost functions c_1 and c_2 so that we are charged $c_1(\theta)$ per subject recruited and $c_2(\theta)$ per unit of time which elapses until a decision is made. Minimising G_1, we set $c_1(\delta/2) = a/n_{fix}$, $c_2(\delta/2) = b/t_{fix}$ and $c_1(\theta) = c_2(\theta) = 0$ otherwise. In addition, we place a uniform three-point prior on θ with $\pi(0) = \pi(\delta/2) = \pi(\delta) = 1/3$. Upon making decision A, if the true effect size is θ, we incur a loss $L(A, \theta)$. For a given pair of decision costs $d_0 > 0$ and $d_1 > 0$, we define our decision loss function such that $L(\text{Accept } H_0, \delta) = d_0$ and $L(\text{Reject } H_0, 0) = d_1$. Then, the total expected cost of the

trial is given by

$$1/3\,\{d_1\,\mathbb{P}(\text{Accept } H_1|\theta=0)+d_0\,\mathbb{P}(\text{Accept } H_0|\theta=\delta)+G_1\}. \tag{8.2}$$

We will argue later that the Bayes problem defined by the pair of decision costs (d_0, d_1) will have a unique solution up to sets of measure zero. Let $\alpha^\star$ and $\beta^\star$ denote the type I rate at $\theta = 0$ and type II error rate at $\theta = \delta$ of the Bayes test minimising (8.2). It follows from the usual Lagrangian argument that in the class of tests with error probabilities $\alpha^\star$ and $\beta^\star$, the Bayes test minimising (8.2) will also minimise G_1. However, in general, $\alpha^\star \neq \alpha$ and $\beta^\star \neq \beta$. To find the solution to our original frequentist problem, we must conduct a two-dimensional search over possible pairs of log decision costs to find the pair $(d_0^\star, d_1^\star)$ defining a Bayes problem whose (unique) solution has $\alpha^\star = \alpha$ and $\beta^\star = \beta$. Since this solution is unique up to sets of measure zero, it must follow that this is also the unique solution to our original frequentist problem we sought to solve.

In the next section we explain how one can find the solution to the Bayes problem defined by a given pair of decision costs (d_0, d_1) using the technique of backwards induction.

8.2.2 Backwards induction

The decision rule of the optimal test minimising (8.2) stipulates that at each stage we proceed by taking the action associated with the minimum expected additional cost. For each $k = 1, \ldots, K$, at decision analysis k the test must either reject or accept $H_0 : \theta \leq 0$. The optimal decision rule can be found analytically, without the use of backwards induction. To see this, let $\tilde{\pi}^{(k)}(\theta|\tilde{z}_k)$ denote the posterior distribution for θ given $\tilde{Z}_k = \tilde{z}_k$. The minimum expected additional cost associated with stopping and making a decision is

$$\eta^{(k)}(\tilde{z}_k) = \min\{\, d_0\tilde{\pi}^{(k)}(\delta|\tilde{z}_k),\, d_1\tilde{\pi}^{(k)}(0|\tilde{z}_k)\,\},$$

achieved by stopping to reject H_0 when $\tilde{z}_k > c_k$ and accepting H_0 when $\tilde{z}_k < c_k$, where c_k is given by

$$c_k = \frac{1}{\delta\sqrt{\tilde{I}_k}}\log(d_1/d_0) + \frac{\delta\sqrt{\tilde{I}_k}}{2},$$

At $\tilde{z}_k = c_k$, the expected additional costs from making either decision are equal.

At interim analysis k, where $k \in \{1, \ldots, K-1\}$, we can take $(K-k)$ more groups of observations, meaning that the optimal test can proceed in one of three ways:

1. Terminate immediately and make the decision, i.e., reject or accept H_0, associated with the minimal additional expected cost. We refer to this as action (1).
2. Halt recruitment and wait for all current pipeline information to become available before making an optimal decision. We refer to this as action (2).
3. Continue to interim analysis $(k+1)$ and proceed optimally thereafter. We refer to this as action (3).

At interim analysis K, only actions (1) and (2) are available; continued sampling is not permitted as the maximum sample size has been reached. Clearly, the expected additional costs associated with actions (2) and (3) will depend on how the test will proceed in later stages. Hence, we must use the technique of backwards induction and find the optimal decision rule at each interim analysis starting at stage K.

For each $k = 1, \ldots, K$, let $\pi^{(k)}(\theta|z_k)$ denote the posterior for θ upon observing $Z_k = z_k$ at interim analysis k. The expected additional cost associated with action (1) is

$$\gamma^{(k)}(z_k) = \min\{\, d_0\pi^{(k)}(\delta|z_k),\, d_1\pi^{(k)}(0|z_k)\,\},$$

achieved by rejecting H_0 for

$$z_k \geq \frac{1}{\delta\sqrt{I_k}}\log(d_1/d_0) + \frac{\delta\sqrt{I_k}}{2}$$

and accepting H_0 otherwise. Let $g_k(\tilde{z}_k|z_k)$ denote the conditional density of $\tilde{Z}_k$ at $\tilde{z}_k$ given $Z_k = z_k$, where g_k is a mixture of normals based on the posterior $\pi^{(k)}$. The expected additional cost associated with action (2) is

$$\rho^{(k)}(z_k) = \frac{b}{t_{fix}}(\tilde{t}_k - t_k)\pi^{(k)}(\delta/2|z_k) + \int_{\tilde{z}_k} \eta^{(k)}(\tilde{z}_k)g_k(\tilde{z}_k|z_k)d\tilde{z}_k. \tag{8.3}$$

The integral on the RHS of (8.3) can be written as

$$d_1\,\pi^{(k)}(0|z_k)\mathbb{P}(\tilde{Z}_k \geq c_k|z_k, \theta = 0) \;+\; d_0\,\pi^{(k)}(\delta|z_k)\mathbb{P}(\tilde{Z}_k < c_k|z_k, \theta = \delta).$$

and hence can be calculated numerically by calling library routines for evaluating probabilities for a standard normal variate.

For $k = 1, \ldots, K-2$, the expected additional cost at interim analysis k associated with

continuing to stage $(k+1)$ and proceeding optimally thereafter is

$$\beta^{(k)}(z_k) = \pi^{(k)}(\delta/2|z_k)\left\{\frac{a}{n_{fix}}(\tilde{n}_{k+1}-\tilde{n}_k)+\frac{b}{t_{fix}}(t_{k+1}-t_k)\right\} + \int_{z_{k+1}} \min\{\gamma^{(k+1)}(z_{k+1}),\rho^{(k+1)}(z_{k+1}),\beta^{(k+1)}(z_{k+1})\}f_{k+1}(z_{k+1}|z_k)\,dz_{k+1},$$

where $f_{k+1}(z_{k+1}|z_k)$ is the conditional density of Z_{k+1} at z_{k+1} given $Z_k = z_k$, found as a mixture of normal densities based on the posterior distribution $\pi^{(k)}$. At interim analysis $K-1$, when only one more group of subjects can be accrued, the expected additional cost of continuing takes a different form; recruitment must close once this group has been accrued so that only actions (1) and (2) are permitted at interim analysis K. Therefore, at interim analysis $K-1$, the expected additional cost associated with action (3) simplifies to

$$\beta^{(K-1)}(z_{K-1}) = \pi^{(K-1)}(\delta/2|z_{K-1})\left\{\frac{a}{n_{fix}}(n_{max}-\tilde{n}_{K-1})+\frac{b}{t_{fix}}(t_K-t_{K-1})\right\} + \int_{z_K} \min\{\gamma^{(K)}(z_K),\rho^{(K)}(z_K)\}f_K(z_K|z_{K-1})\,dz_K. \tag{8.4}$$

Starting at stage K and working backwards, the functions $\beta^{(k)}(z_k)$ are computed iteratively and compared at each stage with $\rho^{(k)}(z_k)$ and $\gamma^{(k)}(z_k)$. The critical values defining the optimal stopping rule are then found as the endpoints of intervals occupied by optimal actions. In the next section, we shall describe how to implement the backwards induction technique described above to find the optimal group sequential test we seek.

8.2.3 Implementation

For a given pair of decision costs (d_0, d_1), we find a solution to the Bayes problem they define starting at interim analysis K. At this stage, our GST can proceed in one of two ways: we close recruitment and either terminate immediately or wait for the pipeline subjects. Optimal actions will occupy intervals. Critical values defining the optimal stopping rule are found as the endpoints of these intervals, i.e., we want to find solutions to the equation $\gamma^{(K)}(z_K) = \rho^{(K)}(z_K)$. To do this, we create a grid of n values of z_K, denoted $z_K[1],\ldots,z_K[n]$, which is efficient for integrating the marginal density of Z_K given our prior for θ. We move through this mesh evaluating $\gamma^{(K)}(z_K) - \rho^{(K)}(z_K)$, looking for changes in sign of this difference between successive grid points which indicate that a change in optimal action has occurred. Suppose we find this has happened between grid points $z_K[\,i\,]$ and $z_K[\,i+1\,]$. A bisection search, followed by linear interpolation, is then used to find the critical value lying in this interval to a greater degree of accuracy. Our grid for z_K is then modified to include the critical values

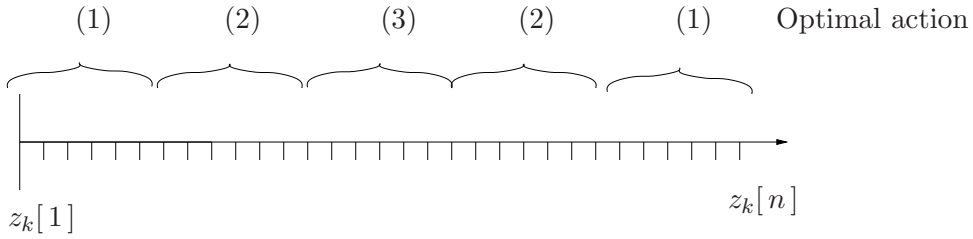


Figure 8-1: Illustration of the optimal action at each point on a grid of values of z_k. Actions (1), (2) and (3) are optimal on intervals which are regarded to be in a "standard" order for tests of the expected form.

found by this search. The expected additional cost of the optimal action at each point in our new grid is computed and stored, ready to be used to calculate $\beta^{(K-1)}(z_{K-1})$ at a grid of values of z_{K-1}.

At interim analysis $K-1$, there are three possible actions. To find the endpoints of intervals of values of z_{K-1} on which each action is optimal we proceed by evaluating the expected additional costs of actions (1) to (3) for a grid of values of z_{K-1}, recording for each grid point the action achieving the minimum. If the optimal action at two successive grid points differs, we conclude the cost curves for these two actions have crossed in this interval. For example, let action (1) be optimal at grid point $z_{K-1}[\,i\,]$ and action (2) be optimal at $z_{K-1}[\,i+1\,]$. Then, we use a bisection search followed by linear interpolation to find the intersection of the curves $\gamma^{(K-1)}(z_{K-1})$ and $\rho^{(K-1)}(z_{K-1})$. Curves $\rho^{(K-1)}(z_{K-1})$ and $\beta^{(K-1)}(z_{K-1})$ would be compared if the optimal action changes from (2) to (3). Figure 8-1 illustrates the order of the intervals that optimal decisions occupy if the optimal test is of the expected form (8.1). Once the boundary constants have been found, we calculate

$$\min\{\gamma^{(K-1)}(z_{K-1}),\, \rho^{(K-1)}(z_{K-1}),\, \beta^{(K-1)}(z_{K-1})\}$$

for a grid of values of z_{K-1} modified to include the critical values found. These minimum costs are then used to evaluate $\beta^{(K-2)}(z_{K-2})$ at the next stage for a grid of values of z_{K-2}. This procedure is repeated until we have found the Bayes test minimising (8.2).

We claim that the optimal test we have found minimising the total expected cost (8.2) is the unique solution, up to sets of Lesbesgue measure zero, to the Bayes problem we set out to solve. To prove this, first suppose the Bayes problem does not have a unique solution. Then, for some k, the cost functions associated with two of the available actions (1) to (3) must be equal over an interval of values of z_k. However, $\gamma^{(k)}(z_k)$, $\rho^{(k)}(z_k)$ and $\beta^{(k)}(z_k)$ are analytic functions of z_k. Hence, following Brown et al. (1980, Theorem 3.3) who reference the arguments of Farrell (1968, Lemma 4.2), we conclude

that these functions are equal either everywhere, which is clearly not the case, or on a set of Lebesgue measure zero. Hence, our claim of uniqueness must follow.

8.2.4 Other objective functions

Tests minimising $G_2, \ldots, G_4$ can be found as the solutions to variants on the Bayes problem formulated and solved in Section 8.2.1. In each instance, all we need do to change is our prior for θ and sampling cost functions c_1 and c_2. For example, looking to minimise G_2, we place a uniform two-point prior on θ, such that $\pi(0) = \pi(\delta) = 1/2$ and $\pi(\theta) = 0$ otherwise. We also set $c_1(0) = c_1(\delta) = a/n_{fix}$, $c_2(0) = c_2(\delta) = b/t_{fix}$ and $c_1(\theta) = c_2(\theta) = 0$ otherwise. Then, the total expected cost of the test is given by

$$1/2\{d_1\mathbb{P}(\text{Accept } H_1|\theta = 0) + d_0\mathbb{P}(\text{Accept } H_0|\theta = \delta)\} + G_2.$$

It follows from the usual arguments that the Bayes test minimising this total expected cost with the required error probabilities will be the optimal frequentist test minimising G_2 that we seek.

Applying the same reasoning to find tests minimising G_3, we set $\pi(-\delta/2) = \pi(0) = \pi(\delta) = \pi(3\delta/2) = 1/4$ and $\pi(\theta) = 0$ otherwise. The sampling cost functions are defined to be $c_1(-\delta/2) = c_1(3\delta/2) = a/n_{fix}$, $c_2(-\delta/2) = c_2(3\delta/2) = b/t_{fix}$ and $c_1(\theta) = c_2(\theta) = 0$ otherwise. Hence, the total expected cost of the test is given by

$$1/4\{d_1\mathbb{P}(\text{Accept } H_1|\theta = 0) + d_0\mathbb{P}(\text{Accept } H_0|\theta = \delta)\} + \frac{1}{2}G_3.$$

Finally, when minimising G_4 we set $\pi(0) = \pi(\delta) = 1/3$ and set a prior probability of $1/3$ on the scenario that $\theta \sim N(\delta/2, (\delta/2)^2)$. The sampling cost functions are defined to be $c_1(\theta) = a/n_{fix}$ and $c_2(\theta) = b/t_{fix}$ if $\theta \notin \{0, \delta\}$ and $c_1(\theta) = c_2(\theta) = 0$ otherwise. Hence,

$$1/3\{d_1\mathbb{P}(\text{Accept } H_1|\theta = 0) + d_0\mathbb{P}(\text{Accept } H_0|\theta = \delta) + G_4\}.$$

In the next section, we present and discuss our findings on the optimal tests found minimising objective functions G_i, $i = 1, \ldots, 4$, for several choices of the weightings a and b.

8.3 Properties of optimal delayed response tests

Tables 8.1 - 8.4 present the minima of G_i, $i = 1, \ldots, 4$, attained by optimal K-stage GSTs when $a = b = 0.5$. Optimality is restricted to the class of tests where the stopping rule at each interim analysis $k = 1, \ldots, K-1$ may be formed of any sequence of

	r						
K	0	0.01	0.1	0.15	0.2	0.25	0.3
2	86.0	88.2	88.6	90.1	91.5	92.9	94.2
3	81.6	81.9	85.5	87.5	89.4	91.2	92.9
5	78.0	78.5	83.0	85.3	87.6	89.8	91.8
10	75.2	75.7	80.9	83.6	86.2	88.5	90.8

Table 8.1: Minima of $100\,G_1$ for tests of $H_0 : \theta \leq 0$ against $H_1 : \theta > 0$ with $\alpha = 0.05$, $\beta = 0.1$, $R = 1.15$, $a = b = 0.5$ and accrual rate $c = 1.0$.

	r						
K	0	0.01	0.1	0.15	0.2	0.25	0.3
2	74.6	74.9	78.1	79.9	81.8	83.6	85.4
3	67.6	68.1	72.8	75.5	78.0	80.5	82.9
5	63.1	63.7	69.4	72.4	75.4	78.2	81.0
10	59.8	60.5	66.8	70.2	73.4	76.5	79.5

Table 8.2: Minima of $100\,G_2$ for tests of $H_0 : \theta \leq 0$ against $H_1 : \theta > 0$ with $\alpha = 0.05$, $\beta = 0.1$, $R = 1.15$, $a = b = 0.5$ and accrual rate $c = 1.0$.

	r						
K	0	0.01	0.1	0.15	0.2	0.25	0.3
2	61.1	61.4	64.5	66.4	68.3	70.2	72.3
3	48.2	48.8	54.6	57.7	60.8	63.8	66.9
5	41.3	42.1	49.3	53.0	56.7	60.3	63.8
10	37.6	38.5	46.2	50.3	54.2	58.0	61.8

Table 8.3: Minima of $100\,G_3$ for tests of $H_0 : \theta \leq 0$ against $H_1 : \theta > 0$ with $\alpha = 0.05$, $\beta = 0.1$, $R = 1.15$, $a = b = 0.5$ and accrual rate $c = 1.0$.

	r						
K	0	0.01	0.1	0.15	0.2	0.25	0.3
2	77.6	77.8	80.7	82.4	84.1	85.7	87.4
3	71.0	71.4	75.8	78.2	80.6	82.9	85.1
5	66.7	67.3	72.6	75.5	78.2	80.8	83.4
10	63.6	64.3	70.3	73.4	76.4	79.2	82.0

Table 8.4: Minima of $100\,G_4$ for delayed response tests for tests of $H_0 : \theta \leq 0$ against $H_1 : \theta > 0$ with $\alpha = 0.05$, $\beta = 0.1$, $R = 1.15$, $a = b = 0.5$ and accrual rate $c = 1.0$.

intervals on which actions (1), (2) or (3) are optimal. For the cases for which we present results, optimal tests are of the form (8.1) we expected. Recall that a and b are the weightings of the averages of $\mathbb{E}(N;\theta)/n_{fix}$ and $\mathbb{E}(T;\theta)/t_{fix}$, respectively. Results are derived under sample size inflation factor $R = 1.15$ and accrual rate $c = 1.0$ so that the

maximum sample size $n_{max} = 1.15\, n_{fix}$ subjects is recruited in $t_{max} = 1.15\, n_{fix}$ units of time. Minima are invariant to changes in the response variance σ^2 and the alternative at which we specify power δ so long as r is equal to the stated value. However, they are not invariant to changes in the accrual rate c, although similar trends have been found to apply under different values of this parameter. Results for $r = 0$ list the minima of G_i when response is immediate. When $c = 1$, the number of subjects recruited at time t is t and so minimising G_i under $a = b = 0.5$ is equivalent to minimising F_i.

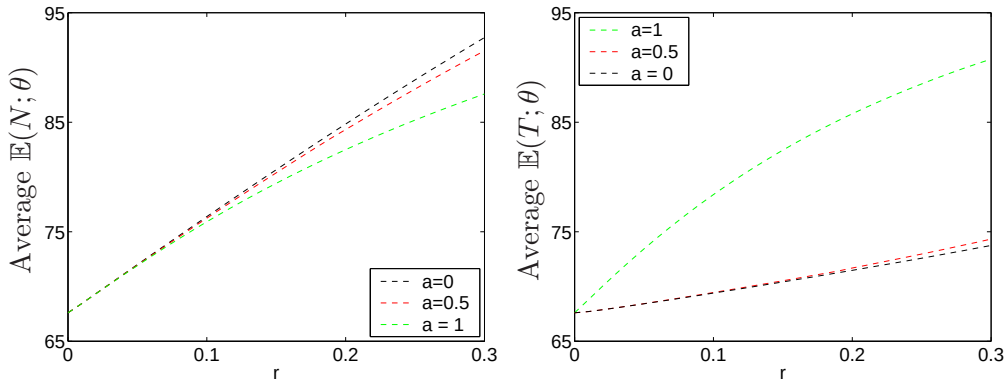
Comparing the minima of $100\, G_i$ with those of F_i presented in Tables 4.2 - 4.5, we find that the efficiency relative to the fixed sample test is now higher. For example, when $r = 0.2$ and $K = 3$, the minimum of $F_2 = 82.5\%$ of n_{fix} and $100\, G_2 = 78.0$. Note that when $r = 0$, $100 G_2 = 67.6$ and we see that even under larger values of r, GSTs minimising G_2 under $a = b = 0.5$ retain many of the benefits for early stopping that can be made when response is immediate. We conclude that if savings in time and sample size are of importance to us, there are certainly worthwhile savings to be made by testing group sequentially for reasonably large values of r.

For $a = 0$ to 1 ($b = 1 - a$), we have calculated the averages of $\mathbb{E}(T;\theta)$ and $\mathbb{E}(N;\theta)$ under $\theta = 0$ and $\theta = \delta$ for designs found minimising

$$G_2 = \sum_{i=0}^{1} \mathbb{E}\left(\frac{aN}{n_{fix}} + \frac{bT}{t_{fix}}; \theta = i\delta\right).$$

If our interest is only in time, the average $\mathbb{E}(N;\theta)$ is minimised when $a = 1$ and $b = 0$. Similarly, the average $\mathbb{E}(T;\theta)$ is minimised when $a = 0$ and $b = 1$. Figures 8-2(a) and 8-2(b) plot the average values of $\mathbb{E}(N;\theta)$ and $\mathbb{E}(T;\theta)$ for these tests optimised for a single objective. When $(a = 0, b = 1)$, optimal tests are highly efficient with respect to both the criteria of minimising $\mathbb{E}(T;\theta)$ and $\mathbb{E}(N;\theta)$. There are good savings to be made on the fixed sample test when we wish for a rapid time to a conclusion. For $r = 0.1$, we can save more than 30% on t_{fix} and for all $r \leq 0.3$, the average $\mathbb{E}(T;\theta)$ stays within 10% of t_{fix} of the average achieved when response is immediate. The average $\mathbb{E}(N;\theta)$ of tests optimal for $(a = 0, b = 1)$ is close to the minima attained when $a = 1$, although for large values of r, the savings on n_{fix} to be made by testing group sequentially are small. Optimal tests for $(a = 1, b = 0)$ perform well only with respect to the criteria for which they have been optimised. Figure 8-2(b) shows that the average $\mathbb{E}(T;\theta)$ of these tests quickly diverges from the minimum as r increases.

Referring to Figures 8-2(a) and 8-2(b), we see that as expected, tests minimising G_2 under $a = b = 0.5$ perform close to optimal with respect to both the criteria of expected sample size and expected time to a conclusion. Their efficiency is also robust to higher



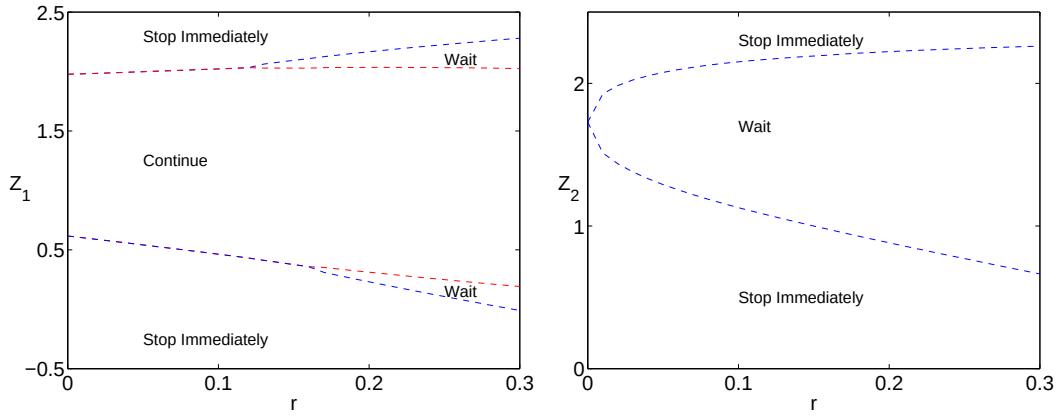
(a) (b)

Figure 8-2: Average values of $\mathbb{E}(N;\theta)$ and $\mathbb{E}(T;\theta)$ under $\theta = 0$ and $\theta = \delta$ expressed as a percentage of n_{fix} and t_{fix}, respectively. Optimal tests minimise G_2 with $K = 3$, $\alpha = 0.05$, $\beta = 0.1$, $R = 1.15$ and accrual rate $c = 1$. Curves appear in the order that they are listed in the legend.

values of r. The average $\mathbb{E}(N;\theta)$ remains within 2% of n_{fix} of the minimum for $r \leq 0.2$, although this gap widens as r increases beyond $r = 0.2$. However, we make large savings on t_{fix} for the average $\mathbb{E}(T;\theta)$ by making the switch from focusing only on sample size to optimising under $a = b = 0.5$. When $r = 0.3$, this saving is more than 15%. This suggests that when our model of the optimisation problem associates costs with accruing subjects and waiting for follow-up, it is sometimes optimal to terminate with an immediate decision at an interim analysis rather than wait for the pipeline information. Figure 8-3(a) shows that this is indeed the case for a two-stage GST minimising G_4 under $a = b = 0.5$. At interim analysis $k = 1$, when r is small there is no interval on which it is optimal to halt recruitment and wait for the pipeline subjects. This is what one would expect based on our results from Section 4.5, where we concluded that even when we are obliged to wait for the pipeline data, in effect the decision of whether to reject or accept H_0 is made at the interim analysis. For larger values of r, we stand to get more out of the pipeline data if we use it. Certainly for the test illustrated in Figure 8-3(a), at interim analysis $k = 1$, the interval on which it is optimal to wait for pipeline data widens. At interim analysis $k = 2$, there are always values of Z_2 for which it is optimal to wait for the decision analysis and this continuation region widens as r increases.

8.4 Optimal tests not of the expected form

For all of the cases for which results are listed in Tables 8.1 - 8.4, optimal tests are of the expected form (8.1) where termination is triggered by larger values of $|Z_k|$. For larger values of r, the optimal tests minimising G_i, $i = 1, \ldots, 4$, are not necessarily of this



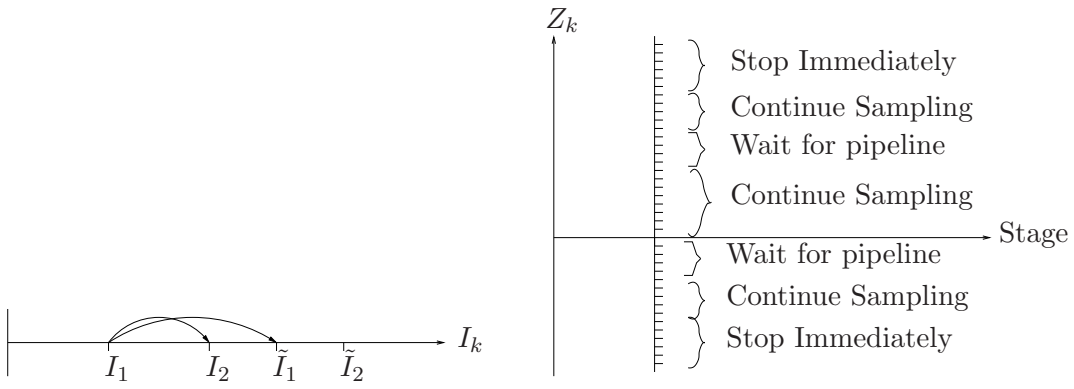
(a) Stopping rule at interim analysis $k = 1$.

(b) Stopping rule at interim analysis $k = 2$.

Figure 8-3: Boundaries of a two-stage one-sided test minimising G_4 for $a = b = 0.5$. Tests are derived under $\alpha = 0.05$, $\beta = 0.1$, $a = b = 0.5$, $R = 1.15$ and accrual rate $c = 1$.

form. Figure 8-4(a) illustrates an example of an information sequence for which optimal decisions do not occupy intervals in the expected order. Instead, as $|z_k|$ becomes large, it is optimal to continue to interim analysis $k + 1$ rather than wait for the pipeline data as expected. In the examples we have considered where the optimal test is not of the form (8.1), we have found that the information sequence and optimal stopping rule are as illustrated in Figures 8-4(a) and 8-4(b). We look to the information sequence to give a possible explanation for the form of the stopping rule. Suppose we observe $Z_k = z_k$ where $|z_k|$ is sufficiently large to motivate closure of recruitment but we wish to postpone making a decision until we have a little more information. However, we don't need all of the information currently in the pipeline. Since for larger values of r, $I_{k+1} < \tilde{I}_k$, the optimal strategy in this case is not to terminate recruitment immediately and wait for the pipeline data, but rather to continue to interim analysis $k+1$ and then terminate with an immediate decision. Hence, the optimal stopping rule is of the form shown in Figure 8-4(b), where intervals on which actions (2) and (1) are optimal are separated by an interval on which it is optimal to continue to interim analysis $(k + 1)$ rather than close recruitment immediately.

At present, our programs are set-up to calculate error probabilities of tests of the form (8.1) but could be easily changed to deal with stopping rules of the type shown in Figure 8-4(b). However, if we wanted to explore this problem further, it would be more sensible to take our procedure even further and allow one to close recruitment at an interim analysis and take any fraction of the pipeline information.



(a) Information sequence when r is large (b) Example of an optimal stopping rule

Figure 8-4: An example of (a) the form of the information sequence generated by a GST when r is large and (b) the stopping rule of an optimal test when r is large.

8.5 Conclusion

This chapter has brought to a conclusion our work on formulating group sequential tests when there is a delay in response. We have extended our original delayed response test structure of (2.2) to allow immediate stopping at an interim analysis. We proceeded to find optimal versions of these tests which minimise objective functions of weighted averages of expected time to a conclusion and sample size. These tests were found to be highly efficient, managing competing aims to achieve a low expected sample size and rapid time to a conclusion. Comparing these tests has also enabled us to further explore and make comments on the influence of the pipeline data on whether we eventually reject or accept H_0.

Up to now, we have assumed there is a non-random delay in response which is equal for each subject. In the next chapter, we extend our treatment of the delayed response problem to explore the problem of designing optimal group sequential survival trials when the delay in response is random and of primary interest.

Chapter 9

Optimal survival trials

9.1 Introduction

In certain disease areas, a pertinent test of a treatment's efficacy is based on analysing the times we wait to observe a particular event. For example, in general, the aim of a Phase III trial in oncology is to compare survival, or progression-free survival, in two treatment groups (Green et al. (2003, p. 8)); times to death from all causes or disease progression are measured relative to some well defined origin, e.g., time of recruitment, coinciding with commencement of treatment. Survival data are sometimes also referred to as time-to-event data.

It is clear that survival data differ from the delayed responses considered earlier in this thesis, for which the delay in response was fixed and uniform across all subjects. Indeed, there are several special features to survival data which render standard methods of statistical inference inappropriate. A histogram of survival times will typically have a heavy upper tail and so it will not be reasonable to assume that they are normally distributed. Survival times may also be censored if, for example, a subject is still alive or yet to relapse, at the time the data are analysed. Most commonly, times are right censored. This occurs repeatedly at the analyses of a GST if a subject is still alive; their true survival time is t, but we have only observed that they have survived to time $c < t$. Right censoring also occurs if subjects are lost to follow-up: for example, they move house and cannot be traced or, if the endpoint is defined as time until death from a particular disease, an individual is judged to have died from an unrelated cause. Left censoring of survival times is less common and occurs when a subject's actual survival time is less than their observed time, i.e., $c > t$. This may be encountered, for instance, when a subject, examined at 2 month intervals for the recurrence of a disease, is first found to have relapsed at 4 months. Then, their relapse-free survival time is

left censored to 4 months. The methods for the analysis of survival data used in this chapter are devised to cope with right censoring.

In a survival study, we accrue information as events, e.g., deaths, relapses or progression of disease, are observed. Testing group sequentially, we use standard designs which stop early for a decision at stage k if our test statistic is sufficiently large or small without waiting to follow-up recruited subjects yet to fail. The first interim analysis must be timed so that the information is not too low. If survival times are short, information will accrue quickly as subjects fail shortly after admittance to the study. Hence, there is the potential for savings in sample size if the trial is terminated at an interim analysis before recruitment has been completed. However, if events are rare and information is slow to accrue, accrual may well be completed before the time of the first interim analysis. Then, the benefits for early stopping of a group sequential design are the potential reductions in trial duration rather than sample size.

In this section, we have outlined some of the issues associated with the analysis of survival data. The particular challenges we face mean that it is not always immediately obvious that designs efficient for more "standard" data, i.e., data following a normal linear model, will also be efficient for survival data. However, in this chapter we present results showing that as far as optimal survival designs go, there is more in common with standard data than one might think; standard GSTs efficient for expected sample size also perform close to optimal for minimising expected time to a conclusion for survival data.

9.2 Methods for analysing survival data

9.2.1 Introduction

Suppose we conduct a trial into the survival of subjects on an experimental treatment. Let T be the non-negative, continuous random variable representing the time until death of an individual which has probability density function $f(t)$. The distribution of T can be summarised by the the survivor function, defined as

$$S(t) = P(T > t) = \int_t^\infty f(u)du, \qquad t \geq 0,$$

and the hazard rate

$$h(t) = \lim_{\delta t \to 0} \left\{ \frac{P(t \leq T < t + \delta t | T \geq t)}{\delta t} \right\} = \frac{f(t)}{S(t)}.$$

We can interpret $h(t)$ as the conditional probability that a subject dies at time t given that they have survived to that time, i.e., it is the instantaneous death rate for subjects surviving to that time. When modelling survival data, one can assume a parametric model for $f(t)$. Green et al. (2003, p. 72) state that it is common to assume survival times are exponentially distributed when determining a clinical trial's sample size requirement. We write $T \sim \text{Exp}(\lambda^{-1})$ and obtain

$$f(t) = \lambda e^{-\lambda t}, \qquad h(t) = \lambda, \qquad \text{for } t \geq 0,$$

so that the hazard of death for an individual after starting treatment remains constant. The exponential distribution has a memoryless property. Hence, given an individual survives to time t, the additional survival time continues to be exponentially distributed with mean λ^{-1}. Alternatively, if the assumption of a constant hazard of death is thought to be unrealistic one can model the data using the Weibull distribution. This distribution is indexed by a scale parameter $\lambda > 0$ and a shape parameter $\gamma > 0$. We write $T \sim W(\lambda, \gamma)$, where

$$f(t) = \lambda\gamma t^{\gamma-1} \exp\{-\lambda t^{\gamma}\} \quad h(t) = \lambda\gamma t^{\gamma-1}.$$

The Weibull distribution was first proposed in the context of industrial testing and is popular in practice because of its flexibility.

9.2.2 Proportional hazards

Suppose we conduct a trial comparing overall survival in two treatment groups labelled A and B, receiving an experimental treatment and control, respectively. The proportional hazards model stipulates that the hazard rate is $h(t)$ for subjects on treatment A and $\lambda h(t)$ for those on treatment B, for $\lambda > 0$. Let $\theta = \log(\lambda)$ denote the log hazard ratio. Should we wish to establish the superiority of treatment A, we test $H_0 : \theta \leq 0$ against $H_1 : \theta > 0$. The assumption of proportional hazards implies that the survivor curves for the two treatments should not cross. However, this will not hold in general. To illustrate this, consider the example given by Whitehead (1997, Section 3.4) of a study comparing a vigorous regimen of chemotherapy against placebo. The demands of the chemotherapy in the short term mean that the failure rate on the active treatment is much higher than on control. However, the chemotherapy has marked long term benefits. Individuals on chemotherapy who survive beyond the first few months go on to have a much higher chance of surviving to five years than those on placebo and the survivor curves for each treatment cross.

We return to our study comparing the survival experiences of subjects taking an

Group	Number of deaths at $t_{i,k}$	Number surviving past $t_{i,k}$	Number at risk just before $t_{i,k}$
A	$r_{iA,k}$	$n_{iA,k} - r_{iA,k}$	$n_{iA,k}$
B	$r_{iB,k}$	$n_{iB,k} - r_{iB,k}$	$n_{iB,k}$
Total	1	$n_{i,k} - 1$	$n_{i,k}$

Table 9.1: The number of deaths and the number surviving beyond the i th ordered death time at interim analysis k.

experimental treatment and control and assume that proportional hazards does hold. We want to design a test of $H_0 : \theta \leq 0$ against $H_1 : \theta > 0$. Often, we do not want to assume a particular functional form for the hazard rate of an individual and so turn instead to non-parametric methods. The log-rank test, proposed by Mantel (1966) and Peto & Peto (1972), is one such procedure which is known to be efficient for detecting a difference in survival times under the proportional hazards model. In the following section, we describe how one can construct group sequential one-sided tests of $H_0 : \theta \leq 0$ based on the log-rank score statistics for testing $H_0 : \theta = 0$. These tests shall then form the basis of our work which shall be presented in subsequent sections.

9.2.3 Log-rank test

Suppose at interim analysis k of our group sequential test of $H_0 : \theta \leq 0$, there are a total of d_k distinct uncensored death times, $t_{1,k} < t_{2,k} < \ldots < t_{d_k,k}$. Let $r_{iA,k}$ and $r_{iB,k}$ denote the numbers of individuals in treatment groups A and B who died at time $t_{i,k}$, respectively. Assume that no death times are tied. Then $r_{iA,k}$ and $r_{iB,k}$, $i = 1, \ldots, d_k$, will be sequences of zeros and ones. Let $n_{iA,k}$ and $n_{iB,k}$ denote the number of subjects alive just before death time $t_{i,k}$ in groups A and B and let $n_{i,k} = n_{iA,k} + n_{iB,k}$. The number at risk includes those who are about to die but excluding those whose right censored time is less than $t_{i,k}$. In the event that a censored survival time is tied with $t_{i,k}$, the values of $n_{iA,k}$ and $n_{iB,k}$ are computed assuming the censored time occurs immediately after the death time. Table 9.1, summarises the situation for the i th ordered death time at interim analysis k in a 2×2 contingency table.

Mantel & Haenszel (1959) proposed the log-rank statistic as a way of combining information over the d_k individual 2×2 tables we shall have at interim analysis k. Fixing the total of each column and row in Table 9.1, we see that the value of $r_{iB,k}$ determines the remaining three entries. Under $\theta = 0$, when there is no difference in the survival experiences on each treatment group, $r_{iB,k}$ has a Bernoulli distribution with expectation $e_{iB,k} = n_{iB,k}/n_{i,k}$. We measure the discrepancy between the observed number of deaths in group B at time $t_{i,k}$ and the number expected under $\theta = 0$ by $r_{iB,k} - e_{iB,k}$. Summing across the d_k failure times at interim analysis k, we obtain the

log-rank score statistic for testing $H_0 : \theta = 0$:

$$S_k = \sum_{i=1}^{d_k} \left\{ r_{iB,k} - \frac{n_{iB,k}}{n_{iA,k} + n_{iB,k}} \right\}.$$

Based on these score statistics, we can define maximum likelihood estimates $\hat{\theta}_k = S_k/I_k$, $k = 1, \ldots, K$. Recall that conditional on $n_{iA,k}$ and $n_{iB,k}$, under $\theta = 0$, $r_{iB,k}$ has a Bernoulli distribution. Then, since all death times are independent, we can approximate the variance of S_k, and hence the information for θ, when $\theta = 0$ by the sum of the conditional variances of $r_{iB,k}$ given $n_{iA,k}$ and $n_{iB,k}$. Doing this, we define

$$I_k = \widehat{var}(S_k) = \sum_{i=1}^{d_k} \frac{n_{iA,k} n_{iB,k}}{(n_{iA,k} + n_{iB,k})^2}, \tag{9.1}$$

Under $\theta = 0$, the survival experiences of subjects on treatments A and B should be the same, i.e., the numbers at risk in both groups should remain approximately equal over time. Substituting this into (9.1), we obtain

$$I_k = \sum_{i=1}^{d_k} \frac{n_{iA,k}^2}{4 n_{iA,k}^2} = \frac{d_k}{4}, \tag{9.2}$$

and we see that information is accrued in a study by observing events. A test's sample size and duration should be chosen to ensure the target number of events is eventually reached.

For θ close to zero, the joint distribution of the sequence $\{S_1, \ldots, S_K\}$ conditional on the observed information sequence $\{I_1, \ldots, I_K\}$ can be approximated by the canonical distribution of (2.5). Harrington et al. (1982) show that this result holds asymptotically, while the small sample accuracy of the approximation has been demonstrated via simulation by, amongst others, Jennison & Turnbull (1984). This result means that we can formulate group sequential designs for survival data based on $S_1, \ldots, S_K$ as we would for "standard data", i.e., data following a normal linear model, when response is immediate. For example, standard two-sided error spending tests can be used to test $H_0 : \theta = 0$ against a two-sided alternative while controlling the type I error rate at its nominal value under any sequence of observed information levels. Likewise, a standard one-sided error spending test could be used to test $H_0 : \theta \leq 0$ against $H_1 : \theta > 0$.

9.3 Formulation of the optimisation problem

We continue with the comparison of the survival experiences of individuals in treatment groups A and B who receive a new treatment and control, respectively. The outcome of interest is the length of time from admittance to the study, which is assumed to coincide with start of treatment, until death from all causes. Let $T_{A,i} \sim \text{Exp}(\lambda_A^{-1})$ and $T_{B,i} \sim \text{Exp}(\lambda_B^{-1})$, $i = 1, 2, \ldots$, represent the survival times of subjects on the new treatment and control, respectively. All failure times are assumed to be independent. Under this exponential model for the data, the hazard rates of death for individuals on each treatment are proportional; from the time of entry into the study, there is a constant hazard of death of λ_A for individuals on the new treatment and λ_B for those on control. Define $\theta = \log(\lambda_B/\lambda_A)$ to be the log hazard ratio for the new treatment against control.

We wish to design a K-stage GST based on the log-rank score statistic for testing $H_0 : \theta \leq 0$ against $H_1 : \theta > 0$ with type I error probability α at $\theta = 0$ and power $1 - \beta$ at $\theta = \delta$. As mentioned in the preamble to this chapter, if events and hence information are slow to accrue, recruitment may be completed before the time of the first interim analysis. In this case, the benefits of group sequential testing are the potential savings to be made in trial duration rather than sample size. Hence, we seek tests optimal for expected time to a conclusion which stop at stage k for rejection of H_0 if $Z_k \geq u_k$ or acceptance of H_0 if $Z_k \leq l_k$, otherwise we continue to the next analysis. We constrain $u_K = l_K$ to ensure the test terminates properly at the final stage.

Figure 9-1 illustrates the progression of a general survival trial; a period of accrual is followed by a follow-up period of duration F. During a survival trial, it is the observation of events that contributes information. Since clearly one cannot predict exactly how many deaths will occur during a study, there will be some uncertainty about just how many subjects to recruit in order to reach a target maximum information level of I_{max}. For a given accrual rate, shortening the period of accrual and recruiting fewer subjects means the length of follow-up must be increased to ensure we observe the required number of events. The competing aims of recruiting fewer subjects and reducing the test's duration will have to be balanced as they pull in different directions and this trade-off is explored by Kim & Tsiatis (1990). In the following work, we standardise our tests to have an accrual period of unit length. The overall study duration is then $1 + F$.

The information levels that will be observed at each interim analysis of a survival study are unpredictable, even if the exact pattern of recruitment is known, so I_k is a random

Figure 9-1: Illustration of the general form of a survival trial, with a period of accrual followed by a period of follow-up.

variable representing the observed information for θ at time t_k. Our approach within this chapter is to derive optimal tests for the sequence of analysis times $\{t_1, \ldots, t_K\}$ which are associated with the information sequence $\{I_k = kI_{max}/K \; ; \; k = 1, \ldots, K\}$. Each t_k is found as the solution to

$$\mathbb{E}(I_k; \theta = 0) = \frac{k}{K} I_{max} \qquad k = 1, \ldots, K.$$

Hence, in effect we derive optimal tests approximating by assuming observed information is equal to the expected information under $\theta = 0$. This seems sensible since approximating the joint distribution of $(S_1, \ldots, S_K)$ by the canonical distribution (2.5) is only accurate for θ close to zero and sufficiently large I_k. Hence, δ, the alternative at which we specify power, should be close to zero. This implies that a large number of failures will be needed, so that by the law of large numbers, approximating observed information levels by their expected values should be reasonably accurate. Hence, we lose little by deriving expected information levels under $\theta = 0$ as opposed to another of the alternative values of θ we are concerned with. The test's final analysis is scheduled at time t_K when the expected information level has reached our target I_{max}. The probability of an individual dying by time t_k will depend on their time of entry into the study. At each interim analysis, subjects will have been followed-up for different lengths of time depending upon their time of recruitment. Hence, in order to calculate the expected information at time t, we must assume a model for the pattern of recruitment into our study.

Let N_t represent the number of subjects recruited at time t, where $N_0 = 0$. We model the process $\{N_t : t \geq 0\}$ as a Poisson process with intensity $\xi > 0$, in which case we can think of ξ as the subject accrual rate. Under this model, for each t, N_t has a Poisson distribution of parameter ξt and we write $N_t \sim Po(\xi t)$. Under $\theta = 0$, let $\lambda = \lambda_A = \lambda_B$ denote the common hazard rate of death for subjects on treatments A and B. Then, for $i = 1, 2, \ldots$, $T_{A,i}$ and $T_{B,i}$ have an exponential distribution of parameter λ^{-1}. Let the random variable D_t represent the number of deaths by time t. Then,

$$\mathbb{E}(D_s; \theta = 0) = \int_0^{\min\{1,s\}} \xi \, \mathbb{P}(T \leq s - t) \, dt \qquad s > 0, \tag{9.3}$$

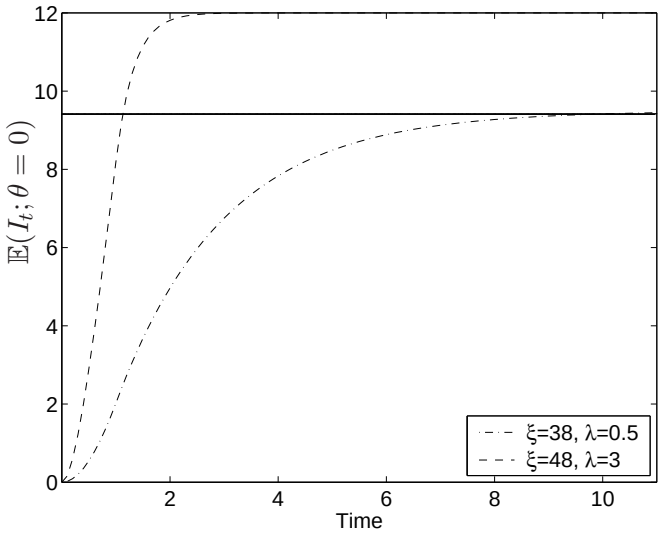


Figure 9-2: Curves of expected information against time for a) $\xi = 38$, $\lambda = 0.5$ and b) $\xi = 48$, $\lambda = 3$. Curves are listed in the order that they appear. Also plotted is the line $1.1\, I_{fix}$ for $\alpha = 0.05$, $\beta = 0.1$ and $\delta = 1$.

where $T \sim \text{Exp}(\lambda^{-1})$. Taking expectations of both sides of (9.2) gives

$$\mathbb{E}(I_k;\theta=0) \approx \frac{1}{4}\mathbb{E}(D_{t_k};\theta=0).$$

Using this approximation, for each $k = 1, \ldots, K$, t_k is found as the solution to

$$\frac{\xi \min\{1,t\}}{4} - \frac{\xi}{4\lambda}\exp(-\lambda t)[\exp(\lambda \min\{1,t\}) - 1] = \frac{k}{K} I_{max}, \tag{9.4}$$

which can be solved using a bisection search over values of t.

In the next sections, we shall present methodology for finding tests of $H_0 : \theta \leq 0$ based on the time and information sequences specified above which are optimal for objective functions concerning time to a conclusion. If there was a linear relation between time and expected information, the GSTs we seek could be found as tests minimising expected sample size when data follow a normal linear model and response is immediate. This problem has been studied by Eales & Jennison (1992) and Barber & Jennison (2002), amongst others. However, Figure 9-2 shows that there is a non-linear relationship between time and expected information for survival data and this facet of our optimisation problem is quite novel.

9.4 Optimal group sequential tests for survival data

Recall that $\theta = \log(\lambda_B/\lambda_A)$ is the log hazard ratio for a new treatment compared against control. We want to design a K-stage GST of $H_0 : \theta \leq 0$ against $H_1 : \theta > 0$ based on the log-rank score statistic with type I error probability α at $\theta = 0$ and power $1 - \beta$ at $\theta = \delta$. The test's maximum information level is set to $I_{max} = RI_{fix}$. Let the random variable W represent the time until termination of the study, either for rejection or acceptance of H_0. Our goal in this section is to find tests of $H_0 : \theta \leq 0$ with the required error probabilities minimising objective functions

$$G_1 = \mathbb{E}(W; \theta = \delta/2),$$

$$G_2 = 0.5\{\mathbb{E}(W; \theta = 0) + \mathbb{E}(W; \theta = \delta)\},$$

$$G_3 = 0.5\{\mathbb{E}(W; \theta = -\delta/2) + \mathbb{E}(W; \theta = 3\delta/2)\},$$

$$G_4 = \int \mathbb{E}(W; \theta) \frac{2}{\delta} \phi\left(\frac{\theta - \delta/2}{\delta/2}\right) d\theta.$$

Note that $G_1, \ldots, G_4$ are analogues of $F_1, \ldots, F_4$, defined in Section 3.1 as objective functions of expected sample size. As outlined in Section 9.3, tests are designed modelling accrual as a Poisson process with intensity ξ. Let λ denote the underlying failure rate under $\theta = 0$, when the survival experiences of individuals in both treatment groups are the same. Given ξ and λ, for each $k = 1, \ldots, K$, analysis time t_k is found as the solution to (9.4). Tests are designed for the information sequence $\{I_k = kI_{max}/K;\ k = 1, \ldots, K\}$ assuming there will be no variation about these values upon their implementation. We approximate the conditional joint distribution of the log-rank score statistics $S_1, \ldots, S_K$ given this information sequence by the canonical distribution (2.5). Define standardised test statistics $Z_k = S_k/\sqrt{I_k}$, $k = 1, \ldots, K$.

For each G_i, we find the optimal test with the required error probabilities by searching for the unconstrained Bayes problem whose unique solution is the group sequential test we seek. The backwards induction algorithm used to solve each of these Bayes problems is very similar to that given by Barber & Jennison (2002), with the exception that we incur additional sampling costs when we continue to the next analysis for sampling time rather than recruiting subjects. For completeness, however, we proceed to explain in the next section how our optimal tests can be found using backwards induction.

9.4.1 Deriving optimal tests

We illustrate the general methodology used to find optimal tests for each G_i, $i = 1, \ldots, 4$, by first explaining in some detail how to find the test minimising G_1 with type I error rate α at $\theta = 0$ and type II error rate β at $\theta = \delta$.

For objective function G_1, we define a sampling cost function which charges $c(\theta)$ per unit of time that elapses until a decision is made, where $c(\delta/2) = 1$ and $c(\theta) = 0$ otherwise. We set a three-point uniform prior on θ, such that $\pi(0) = \pi(\delta/2) = \pi(\delta) = 1/3$. A loss of $L(\text{Reject } H_0, 0) = d_1$ is incurred upon making a type I error under $\theta = 0$ and a loss of $L(\text{Accept } H_0, \delta) = d_0$ is associated with making a type II error under $\theta = \delta$. Define $L(A, \theta) = 0$ otherwise. The total expected cost of a trial is given by

$$\frac{1}{3}\{G_1 + d_1\mathbb{P}(\text{Reject } H_0|\theta = 0) + d_0\mathbb{P}(\text{Accept } H_0|\theta = \delta)\}. \tag{9.5}$$

For a given pair of decision costs d_0 and d_1, we want to find the Bayes test minimising (9.5) which stops for rejection of H_0 at stage k if $Z_k \geq u_k$ and accepts H_0 if $Z_k \leq l_k$. This can be found using the technique of backwards induction.

Let $\pi^{(k)}(\theta|z_k)$ denote our posterior for θ at stage k given $Z_k = z_k$. At analysis $k = 1, \ldots, K$, the minimum expected additional loss associated with stopping and making a decision is given by

$$\gamma^{(k)}(z_k) = \min\{d_1\pi^{(k)}(0|z_k), d_0\pi^{(k)}(\delta|z_k)\}.$$

This minimum is attained by rejecting H_0 when $Z_k > c_k$ and accepting H_0 when $Z_k < c_k$, where

$$c_k = \frac{1}{\sqrt{I_k}}\left(\frac{\log(d_1/d_0)}{\delta} + \frac{\delta I_k}{2}\right) \quad k = 1, \ldots, K.$$

At $Z_k = c_k$, both decisions are associated with equal additional expected loss. We see that, at analysis K the optimal test will proceed by rejecting H_0 for $Z_K \geq c_K$ and accepting H_0 otherwise, and we set $u_K = l_K = c_K$.

At interim analysis $k = 1, \ldots, K-1$, the optimal Bayes test can proceed in one of two ways: either we stop and make a decision or continue sampling to stage $k+1$ and proceed optimally thereafter. Let $\Delta_k = I_{k+1} - I_k$, $k = 1, \ldots, K$. The conditional density of Z_{k+1} given $Z_k = z_k$ and θ is given by

$$f_{k+1}(z_{k+1}|z_k, \theta) = \frac{\sqrt{I_{k+1}}}{\sqrt{\Delta_{k+1}}}\phi\left(\frac{z_{k+1}\sqrt{I_{k+1}} - z_k\sqrt{I_k} - \theta\Delta_{k+1}}{\sqrt{\Delta_{k+1}}}\right).$$

Then, $g_{k+1}(z_{k+1}|z_k)$, the conditional density of Z_{k+1} at z_{k+1} given $Z_k = z_k$ is found by summing $f_{k+1}(z_{k+1}|z_k, \theta)\,\pi^{(k)}(\theta|z_k)$ over $\theta = 0$, $\delta/2$ and δ. For $k = 1, \ldots, K-2$, given $Z_k = z_k$, the expected additional cost of continuing sampling to stage $k+1$ and proceeding optimally thereafter is

$$\beta^{(k)}(z_k) = (t_{k+1} - t_k)\pi^{(k)}(\delta/2|z_k) + \int \min\{\beta^{(k+1)}(z_{k+1}), \gamma^{(k+1)}(z_{k+1})\} g_{k+1}(z_{k+1}|z_k) dz_{k+1},$$

where $(t_{k+1} - t_k)\pi^{(k)}(\delta/2|z_k)$ is the expected sampling cost of continuing to interim analysis $k+1$. At stage $K-1$,

$$\beta^{(K-1)}(z_{K-1}) = (t_K - t_{K-1})\pi^{(K-1)}(\delta/2|z_{K-1}) + \int \gamma^{(K)}(z_K) g_K(z_K|z_{K-1}) dz_K.$$

Starting at stage $k = K-1$ and working backwards, the critical values l_k and u_k are found as solutions to the equation $\beta^{(k)}(z_k) = \gamma^{(k)}(z_k)$. This equation is solved numerically and we refer to Section 3.3.3 for a complete description of the algorithm used.

For a particular choice of d_0 and d_1, the test minimising (9.5) will have type I error rate $\alpha^\star$ at $\theta = 0$ and type II error rate $\beta^\star$ at $\theta = \delta$. By the usual Lagrangian argument, among all tests with the same error probabilities, the test minimising (9.5) will also minimise G_1. The pair of decision costs $d_0^\star$ and $d_1^\star$ defining the Bayes problem whose solution has error rates $\alpha^\star = \alpha$ and $\beta^\star = \beta$ are then found numerically by minimisation of

$$\{\alpha^\star(\log(d_0), \log(d_1)) - \alpha\}^2 + \{\beta^\star(\log(d_0), \log(d_1)) - \beta\}^2 = 0,$$

which is an unconstrained minimisation problem on $\mathbb{R}^2$.

9.4.2 Other objective functions

For minimising functions $G_2, \ldots, G_4$, we adopt the same general approach as described above but change the definition of the Bayes problem to be solved in each case so that we work under different priors and sampling cost functions. The definition of the decision loss function remains unaltered. Minimising G_2, we set $\pi(0) = \pi(\delta) = 1/2$ and $c(0) = c(\delta) = 1$ and $c(\theta) = 0$ otherwise. In this instance, the total expected cost of the trial is given by

$$\frac{1}{2}\{d_1 \mathbb{P}(\text{Accept } H_1|\theta = 0) + d_0 \mathbb{P}(\text{Accept } H_0|\theta = \delta)\} + G_2.$$

Similarly, for G_3 we set $\pi(-\delta/2) = \pi(0) = \pi(\delta) = \pi(3\delta/2) = 1/4$. We also define $c(-\delta/2) = c(3\delta/2) = 1$, and $c(\theta) = 0$ otherwise. Under these settings, the total expected cost of the trial is given by

$$\frac{1}{4}\{d_1\mathbb{P}(\text{Accept } H_1|\theta = 0) + d_0\mathbb{P}(\text{Accept } H_0|\theta = \delta)\} + \frac{1}{2}G_3.$$

Finally, for objective function G_4, let $\pi(0) = \pi(\delta) = 1/3$ and set a prior probability of $1/3$ on the scenario that $\theta \sim N(\delta/2, (\delta/2)^2)$. The sampling cost function is set to be $c(\delta) = c(0) = 0$, and $c(\theta) = 1$ otherwise. In this case, the total expected cost of the trial is given by

$$\frac{1}{3}\{d_1\mathbb{P}(\text{Accept } H_1|\theta = 0) + d_0\mathbb{P}(\text{Accept } H_0|\theta = \delta) + G_4\}.$$

As is the case with G_1, for each of the objective functions G_i, $i = 2, 3, 4$, the set of recursive relations for finding the critical values by backwards induction can be derived following the same lines of reasoning explained in this section.

9.4.3 An example

Kim & Tsiatis (1990) give an example of a lung cancer study designed to test whether a high-dose regimen of chemotherapy, which we shall refer to as treatment A, is superior to a low-dose, referred to as treatment B. Define $\theta = \log(\lambda_B/\lambda_A)$, where we wish to test $H_0 : \theta \leq 0$ against $H_1 : \theta > 0$ using a $K = 5$-stage GST with type I error rate $\alpha = 0.05$ at $\theta = 0$ and power $1 - \beta = 0.9$ at $\theta = 0.41$. The failure rate on the low dose is known to be $\lambda_B = 1.02$. We deviate from their example slightly and set $I_{max} = 1.15\, I_{fix}$ and model recruitment as a Poisson process with intensity $\xi = 250$. We design for the scenario where information levels $I_k = kI_{max}/5$, for $k = 1, \ldots, 5$, are associated with analysis times $\{t_1 = 0.68, t_2 = 1.00, t_3 = 1.35, t_4 = 1.90, t_5 = 3.26\}$, found by repeatedly solving (9.4) for $k = 1, \ldots, 5$. Let I denote the information on termination of the GST and let $f(\theta)$ denote the density of a normal variate with mean δ and standard deviation $\delta/2$. Optimal tests for our problem are derived which minimise the integral of $\mathbb{E}(W; \theta)$ over $f(\theta)$ and the integral of $\mathbb{E}(I; \theta)$ over $f(\theta)$. Table 9.2 compares the boundaries of these tests. The test optimal for time to a conclusion has a much narrower continuation region at the penultimate analysis, reflecting the fact that the increment $t_5 - t_4$ is much larger than $t_k - t_{k-1}$, for $k = 2, 3, 4$.

	Test minimising $\int \mathbb{E}(W;\theta)f(\theta)d\theta$		Test minimising $\int \mathbb{E}(I;\theta)f(\theta)d\theta$	
k	l_k	u_k	l_k	u_k
1	−1.073	2.954	−0.771	2.648
2	−0.0871	2.462	0.040	2.319
3	0.639	2.148	0.571	2.194
4	1.254	1.886	1.037	2.083
5	1.727	1.727	1.721	1.721

Table 9.2: Boundaries of GSTs of $H_0 : \theta \leq 0$ for survival data when $\alpha = 0.05$, $\beta = 0.1$, $\delta = 0.41$, $R = 1.15$, $\xi = 250$ and $\lambda_B = 1.02$. Boundaries are presented on the standardised test statistic scale.

9.5 Efficiency of error spending tests for survival data

9.5.1 Maximum information error spending designs

Continuing with the problem of testing the equality of hazard rate functions $h_A(t) = \lambda_A$ and $h_B(t) = \lambda_B$, suppose we wish to conduct a K-stage test of $H_0 : \theta \leq 0$ against $H_1 : \theta > 0$ based on the log-rank score statistics with type I error α at $\theta = 0$ and power $1 - \beta$ at $\theta = \delta$. Optimal versions of these tests have been formulated in Section 9.4 for a certain sequence of analysis timings and information levels. When survival trials are conducted however, information will accumulate in unpredictable and often unequal increments. Adopting an error spending approach (Lan & DeMets, 1983) then is a highly attractive option: this ensures that the type I error rate will be controlled at its nominal value under any sequence of information levels. In this section, we consider standard error spending designs rather than the versions for delayed responses that were formulated in Chapter 6.

Maximum information error spending designs stipulate that we set a target I_{max} for the maximum information which is to be accumulated in the absence of early stopping. We also pre-specify two non-decreasing functions f and g satisfying $f(0) = g(0) = 0$, and $f(t) = \alpha$ and $g(t) = \beta$, for $t \geq 1$. When a fraction t of I_{max} has been observed, $f(t)$ and $g(t)$ give the cumulative type I and type II error probabilities to be spent, respectively. At analysis k, information levels $I_1, \ldots, I_k$ will have been observed. Then, critical values l_k and u_k are found as the solutions to

$$\mathbb{P}(l_1 < Z_1 < u_1, \ldots, l_{k-1} < Z_{k-1} < u_{k-1}, Z_k \geq u_k; \theta = 0) = \pi_{1,k},$$
$$\mathbb{P}(l_1 < Z_1 < u_1, \ldots, l_{k-1} < Z_{k-1} < u_{k-1}, Z_k \leq l_k; \theta = \delta) = \pi_{2,k},$$

where $\pi_{1,k} = f(I_k/I_{max}) - f(I_{k-1}/I_{max})$ and $\pi_{2,k} = g(I_k/I_{max}) - g(I_{k-1}/I_{max})$. The error spending test stops for rejection of H_0 if $Z_k \geq u_k$ and acceptance if $Z_k \leq l_k$, otherwise we continue to stage $k + 1$. One simple family of error spending functions is

the ρ family, which is defined and discussed in some detail in Section 6.1.2.

In the following discussion, we refer to as "standard" immediate response data following a normal linear model with known covariance matrix. We assume that equally spaced information levels will be achieved with equally sized groups. This description would cover, for example, independent immediate observations $X_{A,i} \sim N(\mu_A, \sigma^2)$ and $X_{B,i} \sim N(\mu_B, \sigma^2)$, $i = 1, 2, \ldots$, generated by a comparison of treatments A and B, where σ^2 is known and allocation to each treatment is equal. The properties of the ρ-family of error spending tests for such standard data are well known. Barber & Jennison (2002) use optimal versions of standard GSTs to assess the efficiency of ρ-family error spending tests for objective functions $F_1, \ldots, F_4$ concerning $\mathbb{E}(N; \theta)$. Tests are designed and implemented for $K = 10$ equally sized groups generating information levels $I_k = kI_{max}/10$, $k = 1, \ldots, 10$. Error spending tests are found to be highly efficient; for most maximum sample sizes, the values of $F_1, \ldots, F_4$ they achieve are within 5% of n_{fix} of the minimum. However, it is not obvious that this efficiency will be replicated for time to a conclusion for survival data. We have noted that when monitoring survival data, information will not accrue at a constant rate as the trial progresses. The exact relationship will depend on the value of ξ and λ but Figure 9-2 shows that the curve of expected information against time will not be linear in the tails. In this section, we extend our efficiency results for ρ-family tests to assess their performance for $G_1, \ldots, G_4$ when based on accumulating survival data.

Fixing $K = 5$, $\alpha = 0.05$, $\beta = 0.1$ and $\delta = 1$, we consider the efficiency of $\rho = 1$ and $\rho = 2$ error spending tests for various values of the subject accrual rate ξ and λ, the underlying hazard rate for individuals in both treatment groups when $\theta = 0$. Tests are based on equally spaced information levels $\{I_k = kI_{max}/K;\, k = 1, \ldots, K\}$. For $\rho = 2$, error probabilities are spent as a quadratic function of the ratios $I_1/I_{max}, \ldots, I_K/I_{max}$. With equally spaced information levels, there is little chance of stopping at the first analysis. Instead error probabilities are spent rapidly in the final stages; almost two-fifths of α and β are to be spent at the final analysis. Using a bisection search, we find that a target maximum information level of $I_{max} = 1.1\, I_{fix}$ must be reached for the test to terminate properly at the final stage with $u_K = l_K$. Table 9.3 lists the boundary constants for this error spending test and for the $\rho = 1$ test which requires a target maximum information level $I_{max} = 1.257\, I_{fix}$. On a grid of values of (λ, ξ), we compute the analysis timings $t_1, \ldots, t_K$ for a $\rho = 2$ test by solving (9.4) successively for each k; the sequence thus generated will vary as λ and ξ change. We then compute the objective function achieved by the error spending test and express this as a percentage of the minima of G_i attained by the optimal test for this case, denoted $G_i^\star$. These results are plotted in Figure 9-3.

ρ	(l_1, u_1)	(l_2, u_2)	(l_3, u_3)	(l_4, u_4)	u_5
2	$(-1.28, 2.88)$	$(-0.26, 2.47)$	$(0.48, 2.20)$	$(1.11, 1.98)$	1.73
1	$(-0.59, 2.33)$	$(0.16, 2.22)$	$(0.76, 2.12)$	$(1.28, 2.01)$	1.83

Table 9.3: The boundaries of $\rho = 1$ and $\rho = 2$ one-sided error spending tests of $H_0 : \theta \leq 0$ for $K = 5$ analyses at equally spaced information levels. Both tests have type I error $\alpha = 0.05$ and power $1 - \beta = 0.9$ at $\theta = \delta$. Boundaries are presented on the standardised test statistic scale.

Figure 9-3 shows that the error spending test for $\rho = 2$ is highly efficient for survival data. It does surprisingly well on average, performing close to optimal with respect to objective functions G_2 and G_4 on the grid of values for ξ and λ considered. Indeed, for many values of ξ and λ, looking at G_4, the average $\mathbb{E}(W; \theta)$ of the error spending test is within 1% of the optimal test's average $\mathbb{E}(W; \theta)$. The performance of the error spending test with respect to G_3 is not quite so close to optimal, although this is somewhat expected since tests minimising G_3 are optimised under extreme values of θ. Even so, in many cases the value of G_3 achieved by the error spending test is still within 5% of the minimum attained by the optimal test.

Consideration of the relationship between time and expected information as ξ and λ vary is central to explaining why error spending designs should be so efficient for a rapid time to a conclusion for survival data. We have already noted that the curve of expected information versus time will not be linear in the tails. However, when either ξ or λ is sufficiently large, we can rule these parts of the curve out for the purpose of scheduling our test analyses: on the left, it is too early to do an analysis if information is very low; on the right, the target maximum information level is reached before the information starts to level off towards its plateau. Therefore, analyses are scheduled at times on the middle portion of the curve for which there is an approximate linear relation between time and information. As ξ or λ grows large, the accuracy of a linear approximation becomes more accurate and the time increments between analyses $t_{k+1} - t_k$, $k = 1, \ldots, K - 1$, are approximately equal.

To illustrate the pattern of analysis timings, first consider the scheduling of analyses for a 5-stage test. Figure 9-2 plots the curve of expected information against time when we have accrual rate $\xi = 38$ and the failure rate on control is known to be $\lambda_c = 0.5$. Setting $\alpha = 0.05$, $\beta = 0.1$ and $\delta = 1$, the $\rho = 2$ error spending test requires a target maximum information level $I_{max} = 9.4$. For each $k = 1, \ldots, 5$, solving equation (9.4) for t_k generates the sequence of timings $\{t_1 = 1.0, t_2 = 1.5, t_3 = 2.3, t_4 = 3.7, t_5 = 10.1\}$. We see that the curve starts to plateau before the target I_{max} has been reached. Hence, the decision to continue at the 4th interim analysis implicates a large commitment of resources as the study duration is now almost tripled. Compare this to the scheduling of

ρ	$100\,F_1/F_1^\star$	$100F_2/F_2^\star$	$100F_3/F_3^\star$	$100\,F_4/F_4^\star$
2	100.6	101.9	108.4	101.3
1	100.8	100.7	104.3	100.4

Table 9.4: Objective functions $F_1, \ldots, F_4$ achieved by ρ-family error spending tests expressed as a percentage of the minima $F_i^\star$ attained by optimal tests for standard data. All tests of $H_0 : \theta \leq 0$ are designed and implemented for $K = 5$ analyses at equally spaced information levels. All tests have type I error rate $\alpha = 0.05$ at $\theta = 0$ and power $1 - \beta = 0.9$ at $\theta = \delta$.

the analyses if we speed up recruitment to $\xi = 48$ and the hazard rate on control turned out to be much higher at $\lambda_c = 3$. From the expected information curve in Figure 9-2, we see that I_{max} is now reached before the curve begins to plateau. Data are analysed between times $t_1 = 0.4$ and $t_5 = 1.1$. On this interval, a linear approximation to the expected information curve is reasonable accurate. Indeed, we calculate that analyses are to be scheduled at $\{t_1 = 0.4,\, t_2 = 0.6,\, t_3 = 0.8,\, t_4 = 0.9,\, t_5 = 1.1\}$, times which are approximately equally spaced.

Let $F_i^\star$ denote the minima of F_i attained for tests of standard data. Compare Figure 9-3 to the corresponding percentages $100\,F_i/F_i^\star$ listed in Table 9.4 attained by the $\rho = 2$ error spending test for standard data. In line with our reasoning given above, as ξ or λ get large, the error spending test's performance for $\mathbb{E}(W;\theta)$ for survival data approaches its performance for $\mathbb{E}(N;\theta)$ for standard data. It is in the bottom left hand quadrant of the (λ, ξ) grids that larger deviations from these performances are seen. However, in practice we are likely to find ourselves in the top left hand quadrant of the (λ, ξ) grid. After all, it is unlikely that we recruit the bare minimum of subjects and hopefully λ, the failure rate on control, will not be too high.

Consider the $\rho = 1$ error spending test. Under the sequence of equally spaced information levels $\{I_k = kI_{max}/K;\, k = 1, \ldots, K\}$, a target information level $I_{max} = 1.257\, I_{fix}$ is required for the test to terminate correctly at stage $K = 5$ when $\alpha = 0.05$ and $\beta = 0.1$. Setting $\delta = 1$, our target information level is $I_{max} = 10.8$ and for this to be reached in a finite time, $\xi > 43.1$. The critical values for this test are listed in Table 9.3. Using the same procedure as was outlined above for $\rho = 2$, we evaluate the objective functions $G_1, \ldots, G_4$ for the error spending test on a grid of values of λ and ξ. Figure 9-4 plots these values as a percentage of the minima of G_i attained for $K = 5$, $\alpha = 0.05$, $\beta = 0.1$, $\delta = 1$ and information inflation factor $R = 1.257$. Looking at the lower left hand quadrant of the graphs, for objective functions G_1, G_2 and G_4, the efficiency of the $\rho = 1$ error spending test is not so robust as that of the test when $\rho = 2$ to lower values of λ and ξ; the higher target information level means that in more cases the expected information curve has started to plateau before I_{max} has been

reached. We conclude that members of the ρ-family of error spending tests designed and implemented under smaller values of ρ and δ will not be so efficient for a rapid time to conclusion for survival data.

9.6 Conclusions

In this chapter, we have derived optimal GSTs for survival data minimising criteria concerning the expected time to a conclusion. Tests are designed and implemented for the idealised scenario where observed information levels are equal to their expected values, although this approximation is fairly accurate. Hence, we claim that our results are still of practical relevance, with the principle "take-home" message being that so far as optimal designs go, tests optimal for "standard" data for the criterion of expected sample size perform close to optimal for expected time to a conclusion for survival data. This efficiency can be explained by consideration of the relationship between expected information and time: the curve is non-linear in the tails but is approximately linear on the time interval over which we are interested in scheduling analyses. Adopting an error spending approach to group sequential testing with survival data is shown to be a pertinent strategy, combining the flexibility to cope with unpredictable information sequences with efficiency. Indeed, for the information and time sequences we have designed for, the values of G_1, G_2 and G_4 achieved by the $\rho = 1$ and $\rho = 2$ error spending tests considered are in many cases within 1% of the minima.

In the next, and final, chapter, we shall summarise the main conclusions of our work and discuss the wider implications for how group sequential tests should be designed when there is a delay in the primary endpoint.

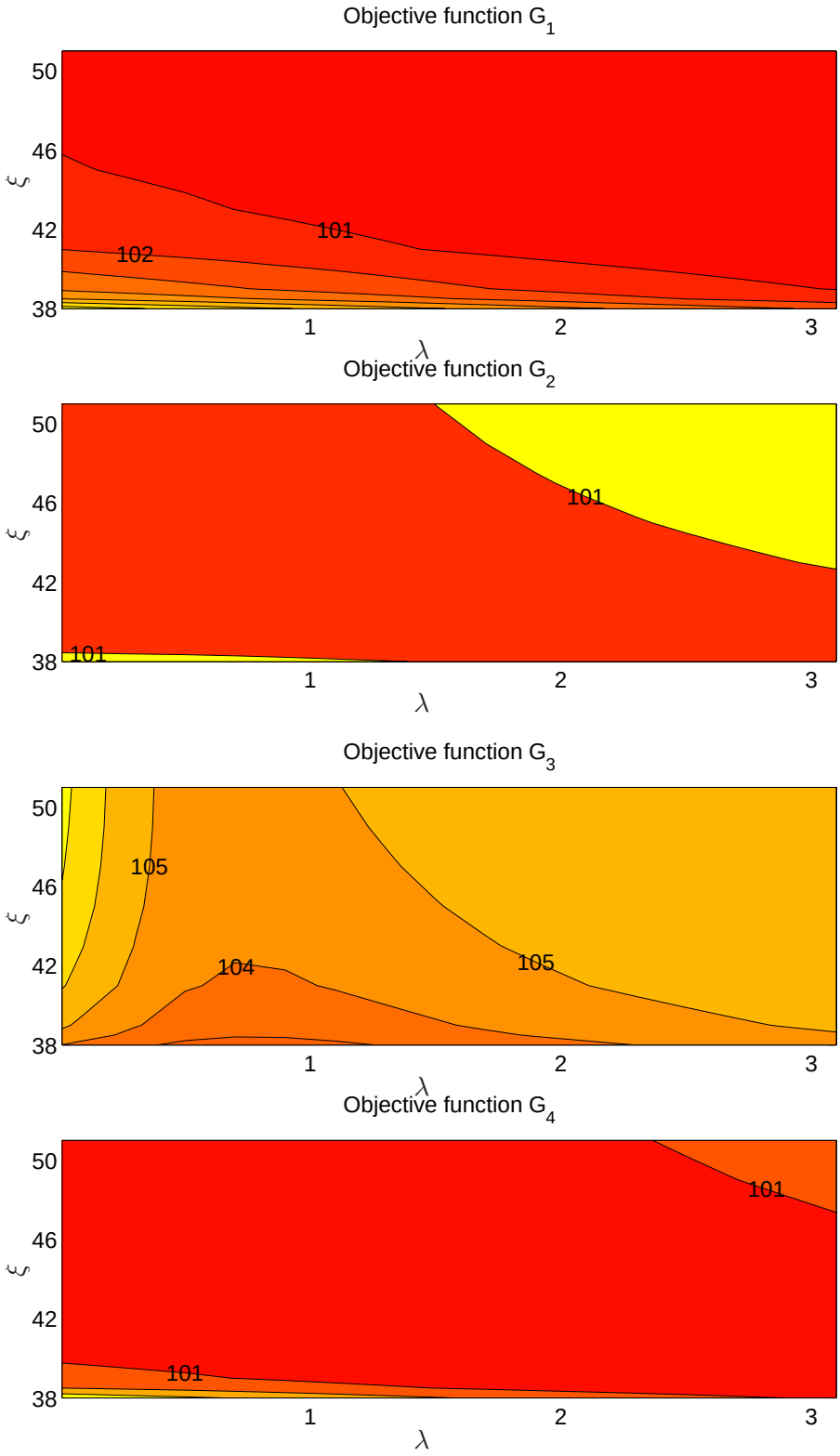


Figure 9-3: Objective functions attained by a $\rho = 2$ error spending test of $H_0 : \theta \leq 0$ expressed as a percentage of the minimum values achieved by optimal tests. Contour lines are spaced at 1% intervals and lighter colours indicate higher percentages. Tests have $\alpha = 0.05$, $\beta = 0.1$, $\delta = 1$, $K = 5$ and target maximum information level $I_{max} = 1.1\, I_{fix}$. All tests are designed under equally spaced information levels.

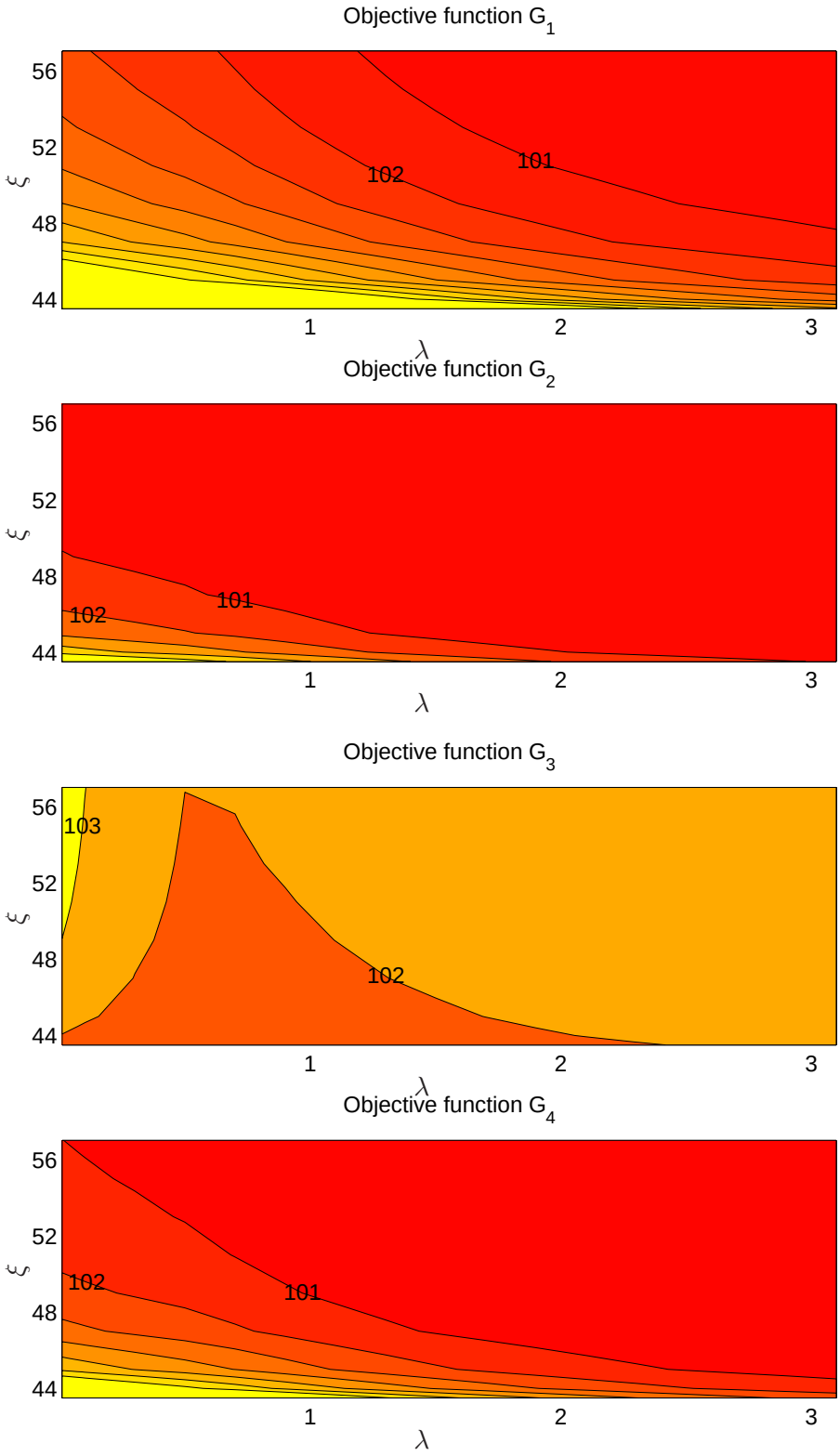


Figure 9-4: Objective functions attained by a $\rho = 1$ error spending test of $H_0 : \theta \leq 0$ expressed as a percentage of the minimum values achieved by optimal tests. Contour lines are spaced at 1% intervals and lighter colours indicate higher percentages. Tests have $\alpha = 0.05$, $\beta = 0.1$, $\delta = 1$, $K = 5$ and target maximum information level $I_{max} = 1.257\, I_{fix}$. All tests are designed under equally spaced information levels.

Chapter 10

Discussion

The benefits for early stopping associated with testing group sequentially when response is immediate are well documented. Optimal versions of standard GSTs have been computed for a variety of criteria involving minimising $\mathbb{E}(N;\theta)$ at several values of θ and savings of over 30% on the fixed sample test reported. Often, however, in practice there will be a delay in the response of clinical interest. In this case, standard GSTs fail to provide a proper treatment of the overrun data that will accumulate after a test's stopping rule has been satisfied. This raises two main questions, namely how should we design GSTs when there is a delay in the response and what, now, are the benefits of group sequential testing for early stopping? In this thesis, we have addressed several facets of the delayed response problem. We have formulated a new class of group sequential designs for delayed responses which provide a proper treatment of any overrun data in a pre-planned way. Optimal versions of these tests have been computed and their properties evaluated. With the formulation of error spending versions of our tests and designs which can incorporate data on a short-term endpoint, our work has culminated in methods which are both user-friendly and flexible enough to be used in practice. Combined with the methods of inference derived in Chapter 5, we now have a complete practical package for dealing with delayed responses.

A natural question is what do we stand to gain from a proper treatment of the pipeline data? Clearly, it ensures the interpretability of a test's results: a delayed response GST gives clear rules, which are stipulated ahead of time, about when to reject H_0 once the pipeline data are in. In terms of efficiency however, the picture is more complex. Certainly, for $r < 0.1$, we don't pay much of a penalty for using a standard GST and ignoring the pipeline data. This is because delayed response GSTs make little use of the pipeline data either and so in both cases, $\mathbb{E}(N;\theta)$ increases by around rn_{max} from their expected sample size when response is immediate. For higher values of r however, delayed response GSTs make greater use of the pipeline data. Tables 4.10 and 4.11 of

Section 4.5 show that these tests terminate recruitment at an interim analysis allowing some probability for "switching" decision once the pipeline data are collected. Hence, for larger values of r, the relative efficiencies of the standard and delayed response tests begin to diverge. Measuring the average $\mathbb{E}(N;\theta)$ under $\theta = 0$ and δ for each test, when $r = 0.2$ the delayed response test saves an additional 3% on n_{fix}. When $r = 0.4$, the additional saving increases to 9%, a substantial saving in the context of the sample size of a Phase III trial.

Evidently, it is only worthwhile switching to our new delayed response designs if a group sequential approach is still viable. Optimal versions of our new tests allow us to assess the savings on the fixed sample test that are possible when there is a delay in response. The benefits of group sequential analysis fall as r increases. However, for smaller values of r, delayed response tests still deliver substantial benefits. Measuring efficiency by criteria involving $\mathbb{E}(N;\theta)$, when $r = 0.1$, a five-stage test still achieves around two-thirds of the savings on the fixed sample test that are made when response is immediate. Even for $r = 0.3$, we still retain around one-third of the savings in sample size. For example, the minimum of the average of $\mathbb{E}(N;\theta)$ under $\theta = 0$ and $\theta = \delta$ is 85.6% of n_{fix}, which amounts to a significant saving in resources in the context of a Phase III study. The benefits of group sequential analysis for a rapid time to a conclusion are more robust to increases in r. For $r \leq 0.3$, tests minimising the average of $\mathbb{E}(W;\theta)$ under $\theta = 0$ and $\theta = \delta$ retain more than 90% of the savings on t_{fix} made when response is immediate.

When the delay in the response of direct clinical interest is long, making measurements on one or more correlated short-term endpoint is an effective way of recouping many of the savings on n_{fix} achieved when response is immediate. This reduces the information in the pipeline at each interim analysis and in effect reduces the value of r we are working under. For example, suppose the delay in a short-term endpoint is one-fifth of that in the primary response and these endpoints have correlation coefficient $\rho = 0.9$. Then, for $r = 0.3$, by making short-term measurements we can save around an additional 10% on n_{fix}, a large gain. In practice, it is unlikely that response variances and correlations will be known exactly. Using error spending versions of our tests, we have formulated a group sequential approach for this case based on the information monitoring strategy of Mehta & Tsiatis (2001) for immediate responses. Hence our designs for incorporating short-term responses offer a solution to the problem of how we can continue to get good benefits for early stopping from group sequential analysis when r is large.

Recall that the delay parameter r is a function of the delay in response and the rate of subject accrual. The fact that the benefits for early stopping decrease with r suggest

that the optimal recruitment strategy when there is a delay in response is not to recruit subjects as quickly as possible. This conclusion echoes comments made by Grieve & Krams (Section 6.1, 2005) who reflect upon a Phase II dose-response trial in stroke following a Bayesian design. Allocation to different doses was adapted as information on dose-response accumulated. Early stopping for futility or success was permitted. The endpoint of direct clinical interest was the Scandinavian Stroke Scale (SSS) 90 days after stroke. For adaptive allocation to be effective, it was essential that recruitment keep pace with learning, otherwise many subjects would have to be randomised before much was known about dose-response and whether the stopping criterion had been met. With this in mind, investigators used SSSs at 1, 4 and 8 weeks, known to be correlated with score at 90 days, to predict a subject's eventual response. Estimates of the dose-response curve could then be updated given these early measurements. In the event, the actual rate of accrual to the study was double that anticipated at the design stage and learning could not take place effectively. Grieve & Krams comment that a slower optimal rate of recruitment would have caused the trial to stop with fewer subjects recruited, a conclusion which concords with our findings.

In the final chapter of this thesis, we turned our attention to survival data, deriving optimal designs for criteria involving $\mathbb{E}(W;\theta)$, approximating observed information by expected information. There are several features to survival data which render usual methods of inference unsuitable. However, we have found that, so far as our optimal designs go, there is more in common with standard data following a normal linear model than one might think. Members of the ρ-family of error spending tests with equally spaced information levels have been investigated. These tests have been found to achieve low $\mathbb{E}(N;\theta)$ for standard immediate response data and we have shown that they also perform close to optimal for $\mathbb{E}(W;\theta)$ for survival data. We conclude that this efficiency is explained by noting that over the time interval where one might be interested in scheduling analyses, expected information increases approximately linearly with time.

We mentioned above that our optimal designs for survival data are derived approximating observed information by expected information. One avenue for further work would be to check the properties of these designs via simulation under several different scenarios. For each simulation replicate, we would proceed by simulating the number of subjects recruited and their arrival times into the study under a Poisson process model for recruitment. Survival times could then be simulated and analysis timings calculated so that observed information levels at each analysis are equally spaced. The actual time until conclusion of the study could then be calculated for this replicate.

Bibliography

Armitage, P. (1958). Numerical studies in the sequential estimation of a binomial parameter. *Biometrika* **45**, 1–15.

Armitage, P., McPherson, C. K. & Rowe, B. C. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society, Series A* **132**, 235–244.

Banerjee, A. & Tsiatis, A. A. (2006). Adaptive two-stage designs in phase II clinical trials. *Statistics in Medicine* **25**, 3382–3395.

Barber, S. & Jennison, C. (2002). Optimal asymmetric one-sided group sequential tests. *Biometrika* **89**, 49–60.

Barnard, G. A. (1946). Sequential tests in industrial statistics. *Journal of the Royal Statistical Society Supplement* **8**, 1–26.

Bauer, P. & Köhne, K. (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics* **50**, 1029–1041.

Bretz, F., Schmidli, H., König, F., Racine, A. & Maurer, W. (2006). Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: General concepts (with discussion). *Biometrical Journal* **48**, 623–634.

Brown, L. D., Cohen, A. & Strawderman, W. E. (1980). Complete classes for sequential tests of hypotheses. *Annals of Statistics* **8**, 377–398.

Chandler, G. A. & Graham, I. G. (1988). The convergence of Nystrom methods for Weiner-Hopf equations. *Numerische Mathematik* **52**, 345–364.

Chang, M. N. (1989). Confidence intervals for a normal mean following a group sequential test. *Biometrics* **45**, 247–254.

Chang, M. N., Hwang, I. K. & Shih, W. J. (1998). Group sequential designs using both type I and type II error probability spending functions. *Communications in Statistics A* **27**, 1323–1339.

Choi, S. C. & Lee, Y. J. (1999). Interim analyses with delayed observations in clinical trials. *Statistics in Medicine* **18**, 1297–1306.

Cui, L., Hung, M. J. & Wang, S. J. (1999). Modification of sample size in group sequential clinical trials. *Biometrics* **55**, 853–857.

Cytel Software Corporation (1992). *EaSt: A software package for the design and analysis of group sequential clinical trials*. Cambridge,Mass.: Cytel Software Corporation.

DeGroot, M. H. (1979). *Optimal Statistical Decisions*. McGraw-Hill.

Denne, J. S. & Jennison, C. (1999). Estimating the sample size for a t-test using an internal pilot. *Statistics in Medicine* **18**, 1575–1585.

Eales, J. & Jennison, C. (1992). An improved method for deriving optimal one-sided group sequential tests. *Biometrika* **79**, 13–24.

EMEA (1998). *E9 Statistical Principles for Clinical Trials*. European Medicines Agency.

Emerson, S. S. & Fleming, T. R. (1990). Parameter estimation following group sequential testing. *Biometrika* **77**, 875–892.

Facey, K. M. (1992). A sequential procedure for a phase II efficacy trial in hypercholesterolemia. *Controlled Clinical Trials* **13**, 122–133.

Fairbanks, K. & Madsen, R. (1982). P-values using a repeated significance test design. *Biometrika* **69**, 69–74.

Faldum, A. & Hommel, G. (2007). Strategies for including patients recruited during interim analysis of clinical trials. *Journal of Biopharmaceutical Statistics* **17**, 1211 – 1225.

Farrell, R. H. (1968). Towards a theory of generalized Bayes tests. *Annals of Mathematical Statistics* **39**, 1–22.

FDA (2004). Innovation/Stagnation: Challenge and opportunity on the critical path to new medical products. Tech. rep., FDA report from March 2004, available at http://www.fda.gov/oc/initiatives/criticalpath/whitepaper.html.

Fisher, L. D. (1998). Self-designing clinical trials. *Statistics in Medicine* **17**, 1551–1562.

Fisher, R. A. (1932). *Statistical Methods for Research Workers*. London: Oliver and Boyd, 4th ed.

Galbraith, S. & Marschner, I. C. (2003). Interim analysis of continuous long-term endpoints in clinical trials with longitudinal outcomes. *Statistics in Medicine* **22**, 1787 –1805.

Ghosh, B. K. (1991). A brief history of sequential analysis. In. *Handbook of Sequential Analysis* , (Eds., B.K. Ghosh and P.K. Sen), New York: Marcel Dekker, 1–19.

Gould, A. L. & Pecore, V. J. (1982). Group sequential methods for clinical trials allowing early acceptance of H_0 and incorporating costs. *Biometrika* **69**, 75–80.

Gould, A. L. & Shih, W. J. (1992). Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance. *Communications in Statistics A* **21**, 2833–2853.

Green, S., Benedetti, J. & Crowley, J. (2003). *Clinical Trials in Oncology*. Boca Raton: Chapman & Hall/CRC, 2nd ed.

Grieve, A. & Krams, M. (2005). ASTIN: A Bayesian adaptive dose-response trial in acute stroke. *Clinical Trials* **2**, 340–351.

Hall, W. J. & Ding, K. (2001). Sequential tests and estimates after overrunning based on p-value combination. Tech. Rep. 01/06, Department of Biostatistics, University of Rochester.

Hall, W. J. & Liu, A. (2002). Sequential tests and estimators after overrunning based on the maximum-likelihood ordering. *Biometrika* **89**, 699–707.

Harrington, D., Fleming, T. & Green, S. (1982). Procedures for serial testing in censored survival data. In. *Survival Analysis* , (Eds., J. Crowley and R.A. Johnson), Hayward, California: Institute of Mathematical Sciences, 269–286.

Haybittle, J. L. (1971). Repeated assessment of results in clinical trials of cancer treatment. *British Journal of Radiology* **44**, 793–797.

Hung, H. M., Wang, S.-J. & O'Neill, R. T. (2006). Methodological issues with adaptation of clinical trial design. *Pharmaceutical Statistics* **5**, 99–107.

Hwang, I. K., Shih, W. J. & DeCani, J. (1990). Group sequential designs using a family of type I error probability spending functions. *Statistics in Medicine* **9**, 1439–1445.

Jennison, C. (1994). Numerical computations for group sequential tests. In. *Computing Science and Statistics* **25**, (Eds., M. Tarter and M. D. Lock), Interface Foundation of America, 263–272.

Jennison, C. (2006). Sample size re-estimation: Internal pilots and information monitoring. Presented at PSI conference, Tortworth Court, Available at http://people.bath.ac.uk/mascj.

Jennison, C. & Turnbull, B. (1984). Repeated confidence intervals for group sequential clinical trials. *Controlled Clinical Trials* **5**, 33–45.

Jennison, C. & Turnbull, B. W. (1997). Group sequential analysis incorporating covariate information. *Journal of the American Statistical Association* **92**, 1330–1341.

Jennison, C. & Turnbull, B. W. (2000). *Group Sequential Methods with Applications to Clinical Trials*. Chapman & Hall/CRC.

Jennison, C. & Turnbull, B. W. (2001). On group sequential tests for data in unequally sized groups and with unknown variance. *Journal of Statistical Planning and Inference* **96**, 263–288.

Jennison, C. & Turnbull, B. W. (2003). Mid-course sample size modification in clinical trials based on the observed treatment effect. *Statistics in Medicine* **22**, 971–993.

Jennison, C. & Turnbull, B. W. (2005). Meta-analyses and adaptive group sequential designs in the clinical development process. *Journal of Biopharmaceutical Statistics* **15**, 537–558.

Jennison, C. & Turnbull, B. W. (2006). Adaptive and nonadaptive group sequential tests. *Biometrika* **93**, 1–21.

Jennison, C. & Turnbull, B. W. (2007). Adaptive seamless designs: Selection and prospective testing of hypotheses. *Journal of Biopharmaceutical Statistics* **17**, 1135–1161.

Keiser, M. & Friede, T. (2000). Re-calculating the sample size in internal pilot study designs with control of the type I error rate. *Statistics in Medicine* **19**, 901–911.

Kiefer, J. & Weiss, L. (1957). Some properties of generalized sequential probability ratio tests. *Annals of Mathematical Statistics* **28**, 57–74.

Kilpatrick, G. S. & Oldham, P. D. (1954). Calcium chloride and adrenaline as bronchial dilators compared by sequential analysis. *British Medical Journal* **ii**, 1388–1391.

Kim, K. & DeMets, D. L. (1987). Confidence intervals following group sequential tests in clinical trials. *Biometrics* **43**, 857–864.

Kim, K. & Tsiatis, A. (1990). Study duration for clinical trials with survival response and early stopping rule. *Biometrics* **46**, 81–92.

Lai, T. L. (1973). Optimal stopping and sequential tests which minimize the maximum sample size. *Annals of Statistics* **1**, 659–673.

Lan, K. K. G. & DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrics* **70**, 659–663.

Lorden, G. (1976). 2-SPRTs and the modified Kiefer-Weiss problem of minimising an expected sample size. *Annals of Statistics* **4**, 281–291.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports* **50**, 163–170.

Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* **22**, 719–748.

Mehta, C. R. & Tsiatis, A. A. (2001). Flexible sample size considerations using information based monitoring. *Drug Information Journal* **35**, 1095–1112.

Müller, H.-H. & Schäfer, H. (2001). Adaptive group sequential designs for clinical trials: Combining the advantages of adaptive and of classical group sequential designs. *Biometrics* **57**, 886–891.

O'Brien, P. C. & Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics* **35**, 549–556.

Pampallona, S. & Tsiatis, A. A. (1994). Group sequential designs for one-sided and two-sided hypothesis testing with provision for early stopping in favour of the null hypothesis. *Journal of Statistical Planning and Inference* **42**, 19–35.

Peto, R. & Peto, J. (1972). Asymptotically efficient rank invariant procedures (with discussion). *Journal of the Royal Statistical Society, Series A* **135**, 185– 206.

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64**, 191–199.

Proschan, M. A. & Hunsberger, S. A. (1995). Designed extensions of studies based on conditional power. *Biometrics* **51**, 1315–1324.

Proschan, M. A., Lan, K. K. G. & Wittes, J. T. (2006). *Statistical Monitoring of Clinical Trials: A Unified Approach*. Springer-Verlag.

Rosner, G. L. & Tsiatis, A. A. (1988). Exact confidence intervals following a group sequential trial: A comparison of methods. *Biometrika* **75**, 723–729.

Schervish, M. J. (1984). Algorithm AS 195: Multivariate normal probabilities with error bound. *Applied Statistics* **33**, 81–94.

Schmitz, N. (1993). *Optimal Sequentially Planned Decision Procedures*, vol. 79 of *Lecture Notes in Statistics*. New York: Springer-Verlag.

Slud, E. V. & Wei, L.-J. (1982). Two-sample repeated significance tests based on the modified Wilcoxon statistic. *Journal of the American Statistical Association* **77**, 862–868.

Sooriyarachchi, M. R., Whitehead, J., Matsushita, T., Bolland, K. & Whitehead, A. (2003). Incorporating data received after a sequential trial has stopped into the final analysis: Implementation and comparison of methods. *Biometrics* **59**, 701–709.

Stallard, N. & Todd, S. (2003). Sequential designs for phase III clinical trials incorporating treatment selection. *Statistics in Medicine* **22**, 689–703.

Timmesfeld, N., Schäfer, H. & Müller, H.-H. (2007). Increasing the sample size during clinical trials with t-distributed test statistics without inflating the type I error rate. *Statistics in Medicine* **26**, 2449–2464.

Todd, S. & Stallard, N. (2005). A new clinical trial design combining phases 2 and 3: Sequential designs with treatment selection and a change of endpoint. *Drug Information Journal* **39**, 109–118.

Tsiatis, A. A., Rosner, G. L. & Mehta, C. R. (1984). Exact confidence intervals following a group sequential test. *Biometrics* **40**, 797–803.

Wald, A. (1947). *Sequential Analysis*. New York: Wiley.

Wald, A. & Wolfowitz, J. (1948). Optimum character of the sequential probability ratio test. *Annals of Mathematical Statistics* **51**, 326–339.

Wang, S. K. & Tsiatis, A. A. (1987). Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics* **43**, 193–200.

Weiss, L. (1962). On sequential tests which minimise the maximum expected sample size. *Journal of the American Statistical Society* **57**, 551–566.

Whitehead, J. (1992). Overrunning and underrunning in sequential clinical trials. *Controlled Clinical Trials* **13**, 106–121.

Whitehead, J. (1997). *The Design and Analysis of Sequential Clinical Trials*. Chichester, U.K.: John Wiley & Sons.

Whitehead, J. & Stratton, I. (1983). Group sequential clinical trials with traingular continuation regions. *Biometrics* **39**, 227–236.

Wittes, J. & Brittain, E. (1990). The role of internal pilot studies in increasing efficiency of clinical trials. *Statistics in Medicine* **9**, 65–72.